Essays

in

Behavioural Financial Markets and Asset Pricing

Dissertation
submitted to the
Faculty of Business, Economics and Informatics
of the University of Zurich

to obtain the degree of
Doktorin der Wirtschaftswissenschaften, Dr. oec.
(corresponds to Doctor of Philosophy, PhD)

presented by

Anastasiia Sokko
from Russian Federation

approved in February 2018 at the request of

Prof. Dr. Thorsten Hens
Prof. Dr. Erich Walter Farkas

The Faculty of Business, Economics and Informatics of the University of Zurich hereby authorizes the printing of this dissertation, without indicating an opinion of the views expressed in the work.

Zurich, 14th February 2018

The Chairman of the Doctoral Board: Prof. Dr. Steven Ongena

Acknowledgements

I would like to take this opportunity and express my sincere appreciation to all who have helped me in any respect during the preparation of this thesis.

First and foremost, I owe my deepest gratitude to my lead supervisor, Prof. Dr. Thorsten Hens, for making it possible for me to work on the topics of my interest in a friendly and flexible environment, for pushing me out of my comfort zone many times, for his patience, motivation, invaluable help and research guidance throughout my PhD study. I also thank Prof. Dr. Thorsten Hens for arranging funding for my doctoral studies. Financial support by the Swiss National Science Foundation - the research grant "Behavioural Financial Markets" - is hereby gratefully acknowledged.

I would like to extend my deepest gratitude to my co-supervisor, Prof. Dr. Erich Walter Farkas, whose research support and reassurance have been of great value to me. Whenever I needed help, he was always there to share his knowledge and guide me.

I am grateful to Prof. Dr. Klaus Reiner Schenk-Hoppé for fascinating discussions, advice on all aspects of the thesis and the invaluable research inspiration he gave me. I am also thankful to his family, Angelika and Linnéa, for making me feel home during my visits in Bremen.

I am thankful to my co-authors Prof. Dr. Ciprian Necula and Dr. Gabriel Drimus for our inspiring and fruitful collaboration. I hope that further projects will follow.

I thank Prof. Dr. Felix Kübler for giving feedback on my research and attending my thesis defence.

Special thanks go to my colleagues and administrative staff of the Department of Banking and Finance and the Chair of Financial Economics in particular, and the S3IT team of the University of Zurich. I thank Egor Maslov, Ferdinand Langnickel, Amelie Brune, Regina Hammerschmid, Daniel Grosshans, Elias Bräm, Alexandra Janssen, Sean Hofland, Richard Meier, Martine Baumgartner, Uschka Dudziak, Antonio Messina, Sergio Maffioletti and many others. I truly appreciate the continuous support, caring assistance and pleasant working environment they provided.

Finally, I dedicate this thesis to my parents for their love, endless support and encouragement.

Zurich, November 2017

Contents

Introduction

Financial markets are an aspect of key importance for creation and protection of financial wealth. It is therefore essential to study and, whenever possible, understand their functioning both on aggregate and individual levels as to enable policy makers as well as individuals to take appropriate and timely preventive measures against adverse market events, such as financial crises and crashes.

Evolutionary finance, which is one of the modern approaches to analyse financial markets, consists in application of Darwinian ideas to asset markets. The approach itself roots back to the 19th century, but has been put in practice only in recent behavioural finance works, e.g., Farmer and Lo (1999), Evstigneev, Hens, and Schenk-Hoppé (2016). It views the market from a biological perspective as a self-organized system of market participants, whose interaction and evolution in the behaviour contributes to a dynamic nature of the market and explains innovations in investment styles, market products as well as changes in the regulatory framework. The Darwinian forces of selection and mutation take effect through the endogenous price process, which determines the evolution of wealth. The adaptive nature of decisions creates a bilateral relation between endogenous asset returns and actions of market participants, and this relation cannot be disentangled.

Classical literature in the area of asset pricing and quantitative finance, on the other hand, analyses markets from an individual investor's perspective. An agent enters into already operating financial market and yet has little or negligible effect on its endogenous dynamics. The price therefore is perceived as an exogenous process by the agent. Two common questions addressed in the literature in this context are whether it is possible, and if so, how to use available price information to evaluate market outcomes, and whether it is possible, and if so, how to use this information to make predictions about current and future market developments.

This dissertation inherits from both classical and evolutionary finance approaches to financial markets and contributes to both strands of the literature.

Chapter 1 introduces institutional details of margin trading into the evolutionary finance model by Evstigneev, Hens, and Schenk-Hoppé (2016) and studies the effects of short selling and leveraging on equilibrium asset prices and market stability. The main finding is that the asset pricing prediction from the evolutionary finance is robust to margin trading, and there is an equilibrium relation between margin requirements and the interest rates on borrowing and lending.

Chapter 2 studies risk-neutral valuation in the context of option pricing. It provides a unified framework for Edgeworth type density expansions (Corrado and Su, 1997, Jarrow and Rudd, 1982, Necula, Drimus, and Farkas, 2016) and derives a closed form option pricing formula applicable to heavy-tailed return distributions characteristic of option markets. The model allows to efficiently recover the risk-neutral density from market quoted option prices. It also provides an alternative to the classical approach of approximating tails of the options implied risk-neutral density by parametric fat-tailed distributions.

Chapter 3 investigates real-time applications of quickest disorder detection techniques to stock market timing. The focus is on the stochastic disorder model and the corresponding exit rule by Zhitlukhin and Ziemba (2016). The paper develops a fully fledged investment strategy that employs signals from the exit rule to enter and exit the market. It finds that the exit rule cannot be employed as a single investment instrument, but acknowledges its risk management abilities, in particular during the fall of 1987 and the bear market of 2007–2009.

References

Corrado, C. J. and T. Su (1997). Implied volatility skews and stock index skewness and kurtosis implied by S&P 500 index option prices. *The Journal of Derivatives 4*(4), 8–19.

Evstigneev, I., T. Hens, and K. R. Schenk-Hoppé (2016). Evolutionary behavioral finance. In *The Handbook of Post Crisis Financial Modeling*, pp. 214–234. Springer.

Farmer, J. D. and A. W. Lo (1999). Frontiers of finance: Evolution and efficient markets. *Proceedings of the National Academy of Sciences 96*(18), 9991–9992.

Jarrow, R. and A. Rudd (1982). Approximate option valuation for arbitrary stochastic processes. *Journal of Financial Economics 10*(3), 347–369.

Necula, C., G. Drimus, and W. Farkas (2016). A general closed form option pricing formula. Working paper.

Zhitlukhin, M. and W. Ziemba (2016). Exit strategies in bubble-like markets using a changepoint model. *Quantitative Finance Letters 4*(1), 47–52.

Chapter I

Margin Requirements and Evolutionary Asset Pricing

joint work with Klaus Reiner Schenk-Hoppé[a]

Abstract. We introduce an evolutionary equilibrium asset pricing model with heterogeneous agents who can either act as brokers or hedge funds. Hedge funds can trade on margin, taking short or (leveraged) long positions in the assets. Brokers provide asset loans and credit to margin traders. In any evolutionary equilibrium, where growth rates of wealth under management are identical, assets are priced according to expected relative dividends (the Kelly rule) and margin traders either leverage long or short the Kelly portfolio. Margin requirements affect the equilibrium interest rates but not the level of asset prices. We also apply the model to study the impact of margin requirements on the speed of price adjustment in the presence of noise traders.

Keywords: Margin trading; short selling; brokers; evolutionary finance.
JEL classification: G11; G12.

[a]Economics, School of Social Sciences, University of Manchester, United Kingdom, and Department of Finance, NHH–Norwegian School of Economics, Norway

1 Introduction

The paper introduces margin trading into the evolutionary finance model by Evstigneev, Hens, and Schenk-Hoppé (2016, 2006, 2008, 2009, 2011) to explore the effect of short selling and leveraging on asset prices and market stability. The present model captures the main institutional detail of margin trading by modelling the relationship between margin traders (hedge funds) and brokers.

Hedge funds can sell short as well as leverage long. But they can only access the market through a broker who executes orders on their behalf and enforces compliance with margin requirements which determine the maximum size of a short or a leveraged position relative to the amount of a margin trader's equity. Brokers extend asset loans and credit to margin traders. They also manage other clients' wealth but can invest long only. Brokers and margin traders collect fees proportional to the value of their assets under management.

We show that in any evolutionary equilibrium assets are priced according to expected discounted relative dividends. Margin traders either leverage long or short the market portfolio. An evolutionary equilibrium is characterized by all of the funds delivering the same net returns (after fees). In such a situation, clients have no incentive to change funds and all of the different funds co-exist in the market. The interpretation of this equilibrium is that it is the outcome of the process of market selection. Market selection refers to underperforming funds losing wealth (and clients) until they have no impact on the market. In an evolutionary equilibrium therefore all of the funds have market impact and their investment decisions matter.

Our results on equilibrium asset prices are in line with previous work in evolutionary finance. The interesting observation here is that margin trading, through which traders can amplify their impact on the market, does not have an effect on equilibrium prices. In the paper by Gerber, Hens, and Woehrmann (2010), which gives a similar extension of the evolutionary finance model but without any institutional detail, short-selling is allowed but not practiced in equilibrium.[1] But in our paper, short selling and leveraging do occur and their extent changes with margin requirements.

The reason for margin trading having no short- or long-term impact on equilibrium prices lies in the institutional structure. In the short term, margin traders' demand resp. supply is offset by brokers because their equilibrium strategies are the same: Leveraged long margin traders borrow money from a broker. This increases the margin traders' short-term demand for the asset but decreases that of the broker. These two effects offset each other. Short sellers increase short-term asset supply but have to deposit their short sale proceeds with the broker. This increases the broker's funds and therefore demand. Again this entails no net effect

[1] Although the absence of short-selling in equilibrium is not explicitly stated in Gerber et al. (2010), this follows from the formula for λ^i_{t0} in their Theorem 1 and the fact that $0 < \beta_i < 1$.

in the short-term. This mechanism is independent of the margin requirement.

Long-term effects on prices can therefore occur only if there is a mismatch between the dividend yield of the market portfolio and the cost of trading on margin. This cost is due to brokers charging leveraged long traders for margin loans and paying interest (the so-called rebate rate) to short-sellers for depositing their short sale proceeds. This makes leveraging long more costly and short selling cheaper compared to a scenario without interest rates. Equilibrium interest rates are set such that brokers do not lose or gain money on average. Thus the level of asset prices is independent of margin requirements. But the model implies an equilibrium relationship between margin requirements and the interest rates on borrowing and deposits.

Coen-Pirani (2005) also finds that asset prices are independent of margin requirements but that the interest rate of a risk-free asset is not. In his general equilibrium model, forced sales by leveraged-long margin traders are offset by a lower interest rate which stimulates purchases by more risk-averse investors. The interest rate is higher the higher the (binding) leverage constraint. In our model economic agents have standard CRRA utility rather than recursive preferences. Margin requirements and interest rate for borrowing are negatively related if, for instance, brokers are more patient than margin traders but both have identical risk aversion. In general, the equilibrium relationship is not necessarily monotone and depends on risk aversions and time preferences. Our model suggests that (a) in margin trading the interplay between traders and brokers is important and (b) the equilibrium relationship between margin requirements and the interest rates on borrowing and deposits is intricate.

Asset prices in an evolutionary equilibrium correspond to the Kelly rule[2], and the market portfolio (which is held leveraged long or short by margin traders and long by brokers) is the growth optimum portfolio. Our findings therefore provide an economic equilibrium justification for the use of this portfolio as a benchmark in the mathematical finance literature, e.g., Long (1990), Platen and Heath (2006), MacLean, Thorp, and Ziemba (2010).

We apply the model to study the impact of margin requirements on market stability. To this end, we use the equilibrium investment strategies as primitives in a heterogenous agent-based model (Hommes, 2013, Hommes and Wagener, 2009) corresponding to our setting. Stability is measured by how quickly asset prices converge to fundamental values (Kelly). Asset prices can become dislocated by noise traders as well as mistakes made by margin traders. Our simulation analysis shows that the market corrects mispricing quickly, and the faster, the less stringent are minimum margin requirements. The main mechanism driving this result is the

[2]Assets are priced according to expected discounted relative dividends. If assets were short-lived and Arrow securities, then the price corresponds to the probability of receiving a payoff from that asset. This is the original Kelly rule (Kelly, 1956) for betting at the racetrack.

increase in the relative amount of assets under management of those traders that are overweight in underpriced assets resp. underweight in overpriced assets. Since leverage multiplies the gains and losses associated with wrongly balanced portfolios, the speed of convergence to fundamental values is faster if margin requirements are weaker. However, in the presence of mistakes, long-term volatility is the higher, the more leverage is possible.

This result holds a policy lesson at odds with views expressed by Shiller (2000). If margin trading is mostly used by 'smart money' investors (here, Kelly investors), then weaker margin requirements stabilize markets by correcting major mispricings quicker. Smaller mispricings, on the other hand, are more frequent since noise traders can also leverage more. However, if noise traders dominate the market, then too loose margin requirements lead to market failure and crashes. This situation resembles the account of the 1929 crash as given by Fortune (2000, 2001): Excessive use of debt to buy common stocks led to a boom in stock prices which came to a sudden stop when brokers had to issue margin calls after a dip in the market.

2 The model

We consider an infinite horizon financial economy in which a finite number of heterogeneous agents trade a finite number of long-lived risky assets.

2.1 Assets, agents and investments

Time is discrete, $t = 0, 1, 2, \ldots$. In each period t a state of nature s_t from a finite set $\{1, ..., S\}$ is revealed. The history of states up to period t is $s^t = (s_0, s_1, ..., s_t)$.

Assets. There are K long-lived assets (stocks) that are risky in terms of future dividend payments and prices. There is one unit of each asset and assets are infinitely divisible. The dividend of each asset k is determined by a non-negative (exogenous) process $D_{t,k}(s^t)$. In any period at least one asset pays a strictly positive dividend, i.e., $\sum_{k\in K} D_{t,k}(s^t) > 0$. Dividends are paid in a perishable consumption good (whose price is taken as the numéraire) before trade takes place in a period. Ex-dividend asset prices are denoted by $p_t = (p_{t,1}, ..., p_{t,K})'$. These prices are endogenously determined by short-term equilibrium of asset demand and supply. Wealth can only be transferred across time using the risky assets.

Economic agents are of two types: There are $N \geq 1$ broker-dealers and $M \geq 1$ hedge fund managers (margin traders). Brokers have direct market access but can take non-leveraged long positions only. Margin traders can take both short and (leveraged) long positions but have to use the services of a broker who acts as an intermediary between the traders and the market. $M(i)$ is the set of margin traders with relationship to broker i. We further assume that all margin traders have a relationship to some broker, but each margin trader can only be a client of one broker.

Both brokers and margin traders are fund managers and provide wealth management services to individual investors (which have no market access). These wealth management services are subject to fees. An operation fee rate $f_t^h(s^t) \in (0,1)$ determines the percentage of wealth under management w_t^h withheld by fund $h \in N \cup M$ in period t:

$$\phi_t^h = f_t^h w_t^h.$$

This operation fee is spent by the fund manager exclusively on buying the consumption good. The residual amount,

$$i_t^h = (1 - f_t^h) w_t^h, \tag{2.1}$$

is re-invested by the fund manager on behalf of its clients.

Fund managers maximize discounted expected utilities from consumption over the infinite time horizon. The utility derived from a consumption process $(c_t(s^t)) \geq 0$ is given by

$$U^h(c) = \mathbb{E}_0 \sum_{t=0}^{\infty} \beta^t u^h(c_t(s^t))$$

($\mathbb{E}_0$ denotes expected value at time 0) with $\beta = \beta^h$ the fund manager h's discount factor, $0 < \beta^h < 1$, and the CRRA instantaneous utility u^h with risk aversion η^h[3]

$$u^h(c) = \begin{cases} \frac{1}{1-\eta^h} c^{1-\eta^h}, & \eta^h \in [0,1) \cup (1,\infty), \\ \ln(c), & \eta^h = 1. \end{cases}$$

Investment strategies are described in terms of the proportion of wealth allocated to a particular asset. For margin trader j these proportions are a stochastic process $\mu_t^j(s^t) = (\mu_{t,1}^j(s^t), ..., \mu_{t,K}^j(s^t))$ with

$$0 \leq \sum_{k \in K} |\mu_{t,k}^j| \leq 1/\mathcal{M}. \tag{2.2}$$

$\mathcal{M}$ with $0 < \mathcal{M} \leq 1$ is the margin requirement. His portfolio positions are

$$x_{t,k}^j := i_t^j \frac{\mu_{t,k}^j}{p_{t,k}}, \tag{2.3}$$

where the value $x_{t,k}^j$ is the number of asset k (i_t^j is the total investment of the fund). A trader is long in the asset when $\mu_{t,k}^j \geq 0$ and short when $\mu_{t,k}^j < 0$.[4]

Broker i's investment strategy is a stochastic process $\lambda_t^i(s^t) = (\lambda_{t,1}^i(s^t), ..., \lambda_{t,K}^i(s^t))$,

[3] Agents can differ in risk aversion and time preferences, but they have homogeneous and correct beliefs beliefs about the states of nature (perfect foresight).

[4] Equation (2.2) says that the trader can choose any effective margin (the ratio of equity to the

such that

$$\sum_{k \in K} \lambda_{t,k}^i = 1 \text{ and } \lambda_{t,k}^i \geq 0 \text{ for all } k \in K. \tag{2.4}$$

Brokers' positions are all long and not leveraged.

Margin trading. Brokers establish margin traders' portfolios. To form a position in assets, a margin trader has to open a margin account with a broker and post collateral. These positions are subject to a margin requirement $\mathcal{M}$. The fraction $\mathcal{M}$ is the amount of equity in terms of the percentage of the market value of the position needed to be placed into the margin account prior to the trade. The margin requirement $\mathcal{M}$ determines how much the margin trader can leverage his position: For every \$1 paid into the margin account, the trader can take a position (long or short) worth up to $\$1/\mathcal{M}$.[5]

Buying on margin is the purchase of an asset by paying the initial margin and borrowing the remaining amount from the broker. All assets bought on margin are kept in the account as collateral for the loan. Fully paid and excess margin securities (securities carried out as margin collateral with a value in excess of percent ξ of customers debit balance) are segregated, while the remainder can be lent out by the broker for a short sale.[6] A loan interest rate r_t, charged on debit balances in the margin account, increases the debt of the margin trader to the broker. Asset dividends accrue to the trader.

Selling an asset short means borrowing the asset from the broker and subsequently selling it in the market. Similar to buying on margin, the trader has to deposit an initial margin into the margin account. The short-sale proceeds are not available to the margin trader and retained by the broker as collateral.[7] While the asset is on the loan, the broker pays the short seller interest for the use of collateral at the rebate rate b_t (which accounts for the stock loan fee).[8] The margin trader, in turn, pays the broker all the dividends on the asset. When the borrowed asset is returned to the broker, the short sale position is closed and the trader receives any remaining fund in the margin account.

purchase price of securities) so that

$$m_t^j := 1/\sum_{k \in K} |\mu_{t,k}^j| = i_t^j / \sum_{k \in K} |x_{t,k}^j p_{t,k}| \geq \mathcal{M}.$$

If $m_t^j = \mathcal{M}$, then the margin trader j's portfolio is leveraged to the maximum; if $m_t^j = 1$ the trader is non-leveraged (assets are paid in full); and if $m_t^j = \infty$, the trader deposits funds with a broker but forms no position in the assets.

[5]Respectively, the maximum leverage allowed on the margin account is $1/\mathcal{M}$.

[6]While all margin securities remain on the broker's street name, the broker can lend assets for a short sale only either from his own holdings or non-segragated margin securities of his clients.

[7]Whereas short sale proceeds cannot be used as initial margin for another asset position, together with the corresponding initial margin they are credited to the margin account, thereby reducing the margin loan.

[8]The rebate rate b_t is earned on any positive balance on the margin account; negative values of the rebate rate may occur. Depositing funds with a broker is effectively a risk-free investment which earns the interest rate b_t.

The minimum margin requirement $\mathcal{M}$ (which is constant and identical for leveraged long and short positions) and the segregation ratio ξ are given exogenously to the model. The loan interest rate r_t and the rebate rate b_t are endogenous to the model: while all fund managers are price and interest rate-takers, market competition among brokers results in the most favourable for traders margin loan and rebate rates, determined by the "zero profit" condition (see Definition 3.1 and the discussion after).

Serving margin traders affects brokers by changing the funds available for investment on his own account. Rather than i_t^i (period t wealth under management net operation fees), the broker has

$$i_t^i + \sum_{j \in M(i)} B_t^j$$

with

$$\begin{aligned} B_t^j := &- \sum_{l \in K:\, x_{t,l}^j \geq 0} (1 - m_t^j) x_{t,l}^j p_{t,l} + \sum_{l \in K:\, x_{t,l}^j < 0} (1 + m_t^j)(-x_{t,l}^j) p_{t,l} \\ =& i_t^j \cdot (1 - \sum_{k \in K} \mu_{t,k}^j) \end{aligned} \tag{2.5}$$

the balance in the margin account of trader j (it is negative if the trader uses a margin loan and positive otherwise). Providing margin loans reduces the broker's funds, while serving short sellers increases it through the collateral. Given the clients' positions, the portfolio x_t^i of assets bought by broker i on his own account is determined by his investment strategy λ_t^i:

$$x_{t,k}^i := (i_t^i + \sum_{j \in M(i)} B_t^j) \frac{\lambda_{t,k}^i}{p_{t,k}}. \tag{2.6}$$

The segregation rule prescribes that

$$x_{t,k}^i - \Delta x_{t,k}^i + \sum_{j \in M(i):\, x_{t,k}^j \geq 0} (x_{t,k}^j - \Delta x_{t,k}^j) \geq \sum_{j \in M(i):\, x_{t,k}^j < 0} (-x_{t,k}^j) \tag{2.7}$$

for all $i \in N$ and $k \in K$ such that $\{j \in M(i) :\ x_{t,k}^j < 0\} \neq \emptyset$, where

$$\Delta x_{t,k}^i := \Big[\sum_{j \in M(i)} \sum_{l \in K:\, x_{t,l}^j < 0} (-x_{t,l}^j) p_{t,l} \Big] \frac{\lambda_{t,k}^i}{p_{t,k}}$$

is a part of broker i's portfolio $x_{t,k}^i$ financed by short-selling proceeds and $\Delta x_{t,k}^j \in [0, x_{t,k}^j]$ are the segregated margin securities (see Appendix A for an explicit expression). The term $\Delta x_{t,k}^i$ appears in equation (2.7) because the short-selling proceeds get available to the broker only after the borrowed securities are sold on the market. In presence of short-sellers, condition (2.7) also ensures that each broker has a positive total demand for every asset.

2.2 Dynamics of wealth and asset prices

Wealth dynamics. In the initial period $t = 0$ fund managers are endowed with consumption good (investments of individual clients into the fund). Future wealth under management is determined by the fund manager's operation fee and gains and losses from his investment strategy (margin traders and broker-dealers) and/or from brokerage services (broker-dealers).

A margin trader j with asset portfolio x_t^j and margin account balance B_t^j in period t will have next period wealth:

$$w_{t+1}^j = \sum_{k \in K} x_{t,k}^j (D_{t+1,k} + p_{t+1,k}) + (1 + r_t) \cdot \min(B_t^j, 0) + (1 + b_t) \cdot \max(B_t^j, 0), \quad (2.8)$$

while a broker i's wealth evolves as:

$$w_{t+1}^i = \sum_{k \in K} x_{t,k}^i (D_{t+1,k} + p_{t+1,k}) - \sum_{j \in M(i)} \left[(1 + r_t) \cdot \min(B_t^j, 0) + (1 + b_t) \cdot \max(B_t^j, 0)\right], \quad (2.9)$$

given his portfolio x_t^i and margin account balances B_t^j of clients $j \in M(i)$. A numerical example that illustrates these equations is given in Exhibit 2.1.

Inserting the expressions (2.3), (2.5) and (2.6) for asset portfolios and margin account holdings into the above two equations, we obtain the wealth dynamics[9] (given market prices, dividends and interest rates) of a margin trader as

$$w_{t+1}^j = (1 - f_t^j) w_t^j \cdot R_{t+1}^M(\mu_t^j), \quad (2.10)$$

[9] For the returns on the broker's and margin trader's investment strategies we have:

$$R_{t+1}^B(\lambda_t^i) := \sum_{k \in K} \lambda_{t,k}^i \frac{D_{t+1,k} + p_{t+1,k}}{p_{t,k}},$$

$$R_{t+1}^M(\mu_t^j) := R_{t+1}^B(\mu_t^j) + (1 + r_t) \cdot \min(1 - \overline{\mu}_t^j, 0) + (1 + b_t) \cdot \max\left(1 - \overline{\mu}_t^j, 0\right)$$

with $\overline{\mu}_t^j = \sum_{k \in K} \mu_{t,k}^j$. The term

$$\overline{B_t^i} := \sum_{j \in M(i)} B_t^j$$

is an inflow (outflow) of funds into the broker i's investment strategy from providing brokerage services in period t, while

$$\overline{I_t^i} := \sum_{j \in M(i)} (1 - f_t^j) w_t^j \left[(1 + r_t) \cdot \max\left(\overline{\mu}_t^j - 1, 0\right) + (1 + b_t) \cdot \min\left(\overline{\mu}_t^j - 1, 0\right)\right]$$

corresponds to the subsequent repayment of margin loans and return of collateral at $t + 1$.

and that of a broker as

$$w_{t+1}^i = [(1 - f_t^i)w_t^i + \overline{B_t^i}] \cdot R_{t+1}^B(\lambda_t^i) + \overline{I_t^i}. \tag{2.11}$$

Market clearing. Asset market equilibrium requires asset demand to be equal to supply. The total demand for asset k from broker i in period t is

$$x_{t,k}^i + \sum_{j \in M(i)} x_{t,k}^j \tag{2.12}$$

with $x_{t,k}^i$ assets bought on his own account, $x_{t,k}^j \geq 0$ margin securities bought on behalf of a margin trader j and $-x_{t,k}^j \geq 0$ shorted assets, which are sold back on the market.

Since brokers are the only market participants who actually hold and trade assets and assets are in exogenous unit supply, the market clearing condition in period t is given by:

$$\sum_{j \in M} x_{t,k}^j + \sum_{i \in N} x_{t,k}^i = 1, \quad k \in K. \tag{2.13}$$

Asset price dynamics. The market clearing condition (2.13) determines the market clearing asset prices. By inserting expressions (2.3) and (2.6) for a margin trader's and broker's portfolios and rearranging the terms, one gets:

$$p_{t,k} = \sum_{j \in M} i_t^j \mu_{t,k}^j + \sum_{i \in N} \Big(i_t^i + \sum_{j \in M(i)} i_t^j (1 - \sum_{l \in K} \mu_{t,l}^j) \Big) \lambda_{t,k}^i. \tag{2.14}$$

If there are no margin traders, prices are given by $p_{t,k} = \sum_{i \in N} i_t^i \lambda_{t,k}^i$, as in the standard model.[10] In this model, margin trading impacts asset prices. To quantify the impact of margin traders on asset prices, we consider a small transfer of wealth from a broker to a margin trader. This exercise in comparative statics yields insights on the impact of margin trading.

Assume that at date t a fraction Δw^i of broker i's investment is given to a new margin trader j', who is the broker's client. Then the price of asset k changes by

$$\Delta p_{t,k} = \Delta w^i \mu_{t,k}^{j'} + \Delta w^i \Big(- \sum_{l \in K} \mu_{t,l}^{j'} \Big) \lambda_{t,k}^i.$$

If $\mu_{t,k}^{j'} = 1/m^{j'}$, i.e., the margin trader's portfolio consists only of a leveraged long position in asset k, then the asset price increases by

$$\Delta p_{t,k} = \Delta w^i \frac{1 - \lambda_{t,k}^i}{m^{j'}} \geq 0.$$

[10] In the absence of margin traders, i.e., if $w_0^j = 0$ for $j \in M$, the asset price and wealth dynamics reduces to the one considered in Evstigneev et al. (2006, 2008).

Exhibit 2.1: Margin account of a trader and assets and liabilities of a broker.

Account positions			
stock	qty	price	value
ABC	100	\$ 50	\$ 5,000
XYZ	−50	\$ 10	\$ − 500
Market value		\$ 4,500	
Account balance		\$ − 1,500	
Equity		\$ 3,000	

a) Margin account of trader 1 at t.

Account positions			
stock	qty	price	value
ABC	100	\$ 52	\$ 5,200
XYZ	−50	\$ 11	\$ − 550
Market value		\$ 4,650	
Account balance		\$ − 1,455	
Equity		\$ 3,195	

b) Margin account of trader 1 at $t + 1$.

Account positions			
ABC	−60	\$ 50	\$ − 3,000
XYZ	100	\$ 10	\$ 1,000
Market value		\$ − 2,000	
Account balance		\$ 6,000	
Equity		\$ 4,000	

c) Margin account of trader 2 at t.

Account positions			
ABC	−60	\$ 52	\$ − 3,120
XYZ	100	\$ 11	\$ 1,100
Market value		\$ − 2,020	
Account balance		\$ 6,020	
Equity		\$ 4,000	

d) Margin account of trader 2 at $t + 1$.

Assets			
Receivables			\$ 1,500
Stocks			
ABC	545	\$ 50	\$ 27,250
XYZ	2725	\$ 10	\$ 27,250
Total assets			\$ 56,000
Liabilities			
Payables			\$ 6,000
Total liabilities			\$ 6,000
Equity			\$ 50,000

e) Broker's assets and liabilities at t.

Assets			
Receivables			\$ 1,500
Interest			\$ 30
Stocks			
ABC	545	\$ 52	\$ 28,340
XYZ	2725	\$ 11	\$ 29,975
Dividends			\$ 1,907.5
Total assets			\$ 61,752.5
Liabilities			
Payables			\$ 6,000
Interest			\$ 30
Total liabilities			\$ 6,030
Equity			\$ 55,722.5

f) Broker's assets and liabilities at $t + 1$.

NOTES: Margin accounts of trader 1, trader 2 and assets and liabilities of their broker in periods t and $t + 1$. Dividends $D_{t+1,ABC} = \$\,1$ and $D_{t+1,XYZ} = \$\,0.5$ per unit of stock, loan interest rate $r_t = 2\%$, rebate rate $b_t = 0.5\%$, segregation ratio $\xi = 140\%$.
a) Margin trader 1 invests $i_t^1 = \$\,3,000$ to buy 100 ABC and sell short 50 XYZ (strategy $\mu_t^1 = (5/3, -0.5/3)$, margin equity $m_t^1 = 54.55\%$). 58 excess margin securities ABC segregated, 42 ABC lent out to trader 2 for a short sale. Shorted XYZ borrowed from the broker.
b) Margin loan interest of \$ 30 charged and \$ 100 (\$ 25) dividends from ABC (XYZ) credited (debited) to trader 1's margin account.
c) Margin trader 2 invests $i_t^2 = \$\,4,000$ to buy 100 XYZ and sell short 60 ABC (strategy $\mu_t^2 = (-0.75, 0.25)$, margin equity $m_t^2 = 100\%$). 100 fully paid margin securities XYZ segregated. Shorted ABC borrowed from trader 1 (58 stocks) and broker (2 stocks).
d) Interest of \$ 30 accrued and \$ 50 (\$ 60) dividends from XYZ (ABC) credited (debited) to trader 2's margin account.
e) Broker issues margin loan of \$ 1,500 to finance position of trader 1 and receives \$ 6,000 from trader 2 as collateral. Investment funds \$ 56,000 (portion $i_t^i = \$\,50,000$ of own equity) spent to buy 545 ABC and 2725 XYZ (strategy $\lambda_t = (0.5, 0.5)$). 2 ABC and 50 XYZ lent out to traders 1 and 2 for a short sale.
f) Trader 1 repays broker margin loan \$ 1,500 and pays margin loan interest of \$ 30. Broker returns trader 2 collateral (\$ 6,000) with accrued interest (\$ 30). Stocks ABC and XYZ deliver dividends of \$ 545 and \$ 1,362.5 respectively.

If, conversely, the margin trader sells the asset k short, i.e., $\mu^{j'}_{t,k} = -1/m^{j'}$, then

$$\Delta p_{t,k} = \Delta w^i \frac{\lambda^i_{t,k} - 1}{m^{j'}} \leq 0,$$

i.e., his trade has a negative effect on the asset k's price.

The impact of margin traders on asset prices is realized through the change in the effective asset supply: the left-most term in (2.13) is the adjustment of effective supply of asset k in response to margin trading. Therefore, an increase in leveraged long positions has the same effect as a reduction in supply: it increases the price. Short selling has the opposite effect increasing the supply and, thus, reducing the asset price.

2.3 Interest rates and inflation

Proposition 2.1 below shows that aggregate consumption in any given period equals aggregate dividends paid by risky assets in this period. This implies that if \$1 has a buying power (in terms of consumption good) of $1/\overline{D}_t$ in period t, it would have a buying power of $1/\overline{D}_{t+1}$ in the next period $t+1$ with

$$\overline{D}_t := \sum_{k \in K} D_{t,k}.$$

That means, purchasing power falls at the growth rate of aggregate dividends:

$$g^D_{t+1} := \overline{D}_{t+1}/\overline{D}_t.$$

Proposition 2.1 *In every period $t \geq 0$ the following holds:*

$$\sum_{h \in N \cup M} f^h_t w^h_t = \overline{D}_t. \tag{2.15}$$

Proof of Proposition 2.1: Summing up equations (2.8) and (2.9) over $j \in M$ and $i \in N$ gives

$$\overline{w}_{t+1} = \sum_{k \in K} \sum_{h \in M \cup M} x^h_{t,k}(D_{t+1,k} + p_{t+1,k}) = \overline{D}_{t+1} + \overline{p}_{t+1}$$

with $\overline{w}_{t+1} := \sum_{h \in N \cup M} w^h_{t+1}$ and $\overline{p}_{t+1} := \sum_{k \in K} p_{t+1,k}$, where the last equality follows from the market clearing condition (2.13). Further, summing equations (2.14) over $k \in K$ and using that $\sum_{k \in K} \lambda^i_{t,k} = 1$, we obtain

$$\overline{p}_t = \sum_{h \in N \cup M} i^h_t = \sum_{h \in N \cup M} (1 - f^h_t) w^h_t. \tag{2.16}$$

Inserting this term into the previous equality completes the proof.[11] □

To account for inflation in the cost of borrowing and lending, the following assumption on the interest rates is imposed throughout the remainder of the paper.

Assumption 2.1 *There are constants $r, b > -1$ such that*

$$\frac{1 + r_t(s^{t+1})}{g^D_{t+1}(s^{t+1})} = 1 + r, \quad \frac{1 + b_t(s^{t+1})}{g^D_{t+1}(s^{t+1})} = 1 + b$$

for all $t \geq 0, s^{t+1}$.

The interest rates set by brokers on margin loans and deposits are adjusted to inflation when Assumption 2.1 holds. In real markets these rates are referred to as (short) rebate and margin rate and are typically quoted as (positive or negative) percentage points relative to some base rate such as the Fed Funds rate. In the current model we consider these rates as floating, i.e., when taking a margin loan the exact amount of interest to be paid is not known and benchmarked to the market. Assumption 2.1 implies that the real interest rates (nominal interest rates accounted for inflation) are constant.

3 Evolutionary equilibrium

For given asset prices and interest rates, fund managers choose their investment strategies and operation fees so as to maximize their discounted expected utility from consumption

$$\mathbb{E}_0 \sum_{t=0}^{\infty} \beta^t u(c_t(s^t)). \tag{3.1}$$

[M] Margin trader $j \in M$ maximizes (3.1) with $\beta = \beta^j$ and $u = u^j$ by choosing (f^j, μ^j) subject to

$$0 \leq \sum_{k \in K} |\mu^j_{t,k}| \leq 1/\mathcal{M} \tag{3.2}$$

and

$$c_t = f^j_t w^j_t, \quad 0 < f^j_t(s^t) < 1 \tag{3.3}$$

with

$$w^j_{t+1} = (1 - f^j_t) w^j_t \cdot R^M_{t+1}(\mu^j_t).$$

[B] Broker $i \in N$ maximizes (3.1) with $\beta = \beta^i$ and $u = u^i$ by choosing (f^i, λ^i) subject to

$$\sum_{k \in K} \lambda^i_{t,k} = 1 \text{ and } \lambda^i_{t,k} \geq 0 \text{ for all } k \in K$$

[11] In period $t = 0$ the equality holds by virtue of choice of dividends $\overline{D}_0$. Without loss of generality dividends $\overline{D}_0$ can be set arbitrarily because the payments $D_{0,\cdot}$ enter neither into the wealth dynamics of margin traders (2.8) and brokers (2.9) nor into the market clearing prices (2.14).

and (3.3) with

$$w_{t+1}^i = [(1-f_t^i)w_t^i + \overline{B_t^i}] \cdot R_{t+1}^B(\lambda_t^i) + \overline{I_t^i}.$$

Denote by g_t^h the growth rate of wealth under management of fund h in period t:

$$g_t^h := w_t^h / w_{t-1}^h.$$

The growth rate g_t^h is effectively the net return, which an individual investor receives at date t per \$1 invested in fund h in period $t-1$.

Similar to the Miller and Modigliani (1961)'s "fundamental principle of valuation" approach (under perfect certainty), we consider equilibria with balanced growth rates, i.e., when net returns of all funds are identical. Otherwise, clients of low net return funds would increase their welfare by withdrawing from these funds and investing into ones that offer higher rates of net return. This process would tend to either drive the low net return investment funds out of the market (i.e., discontinue their line of business because of lack of wealth under management) or force them to lower the operation fee rates to compensate for low returns generated by their investment strategies. This motivates the following definition:

Definition 3.1 A collection of operation fee rates $f_t^h(s^t)$, $h \in N \cup M$, investment strategies $\lambda_t^i(s^t)$, $i \in N$, and $\mu_t^j(s^t)$, $j \in M$, together with interest rates r_t and b_t, is an *evolutionary equilibrium*, if

(i) the market clearing prices given by (2.14) are strictly positive, brokers have positive total asset demand (2.12) and the segregation rule (2.7) is satisfied;

(ii) given prices and interest rates, each margin trader's decision (f^j, μ^j) solves the optimization problem **[M]**;

(iii) given prices, interest rates and clients' decisions, each broker's decision (f^i, λ^i) solves the optimization problem **[B]**;

(iv) all economic agents' growth rates of wealth, g_t^h, $h \in N \cup M$, are identical.

The equilibrium is in a *steady state* if operation fee rates and investment strategies are constant.

The property of constant operation fee rates and investment strategies in a steady state means that all investment funds charge a certain (percentage) wealth management fee and hold certain portfolio weights for a long period. If asset prices fluctuate, the fund manager has to adjust the number of shares in the portfolio as to keep the proportions constant. The fee that an individual investor has to pay to his fund manager is proportional to the market value of the purchased portfolio and therefore depends on the overall performance of the fund's investment strategy. In fact, many institutional investors follow these rules, among other things, in order to increase credibility and reduce monitoring costs.

The assumption of constant operation fee rates and investment strategies is neither that restrictive from the general equilibrium point of view as it might seem. Hakansson (1966, 1970), for instance, considers a related infinite horizon portfolio optimization problem with CRRA utilities, borrowing/lending and non-capital income stream but without margin trading. He finds that the optimal consumption (an equivalent of operation fees in the current model) is linear in wealth and the present value of non-capital income stream, while the optimal investment strategies depend on asset returns, being constant if these returns are stationary. Accordingly, as the agents' optimization problems in the present model are, to a certain degree, generalizations of the Hakansson's decision problem, one would expect its equilibria to inherit the property of constant optimal strategies and operation fees rates.[12]

3.1 Kelly in prices and strategies

Let us assume that the margin trader is either long only or short only in all of the assets, i.e., $M = M^L \cup M^S$ with $M^L \cap M^S = \emptyset$ such that $\mu^j_{t,k} \geq 0$ for all $j \in M^L$ and $\mu^j_{t,k} < 0$ for all $j \in M^S$. This assumption simplifies the presentation by avoiding to deal with corner solutions in the first-order conditions which would require distinguishing many different cases. We take a closer look at whether a trader prefers to leverage or to sell short in Section 3.2.1.

We further assume that all margin traders make full use of their margin. Thus, influence of each individual trader's strategy on asset prices would be most pronounced as it is amplified by high leverage used: any change in the level of margin requirements would be directly reflected in equilibrium prices, realized through investment strategies of margin traders (see the pricing equation (2.14)). The more general case when agents' effective margin can be lower than the maximum is considered in the subsequent Section 5.

The following assumption on asset dividends is imposed throughout the remainder of the paper.

Assumption 3.1 *(i) There are no redundant assets, i.e., the $S \times K$ dimensional matrix $D_{t,\cdot}(s^{t-1}, \cdot)$ has rank K;*

(ii) Relative dividend payments have constant expected values:

$$\mathbb{E}_t d_{t+1,k} = \mathbb{E}_t \frac{D_{t+1,k}}{\overline{D}_{t+1}} = d_k$$

with $d_k > 0$ constant, $k \in K$;

[12] Brokerage service and margin trading are effectively borrowing, lending (both funds and stocks) and receiving (possibly negative) non capital income stream in every period.

(iii) The growth rate of aggregate dividends g_t^D is weakly stationary[13] *and bounded, i.e.,*

$$g_t^D(s^t) \leq G \text{ for each } t > 0, s^t$$

with a constant $G < \min_{h \in N \cup M:\, \eta^h < 1} \beta^{1/(\eta-1)} < \infty$;

(iv) $(g_t^D)^{1-\eta^h}$ and d_t are uncorrelated for every $h \in N \cup M$, $t > 0$.

In the case of two risky assets, a dividends process satisfying Assumption 3.1 may be given as follows. Let $\xi_t \in (0,1)$ be an i.i.d. process and $0 < g_t^D \leq G$ another process independent of ξ_t. Then define the aggregate dividends by $\overline{D}_t = \overline{D}_{t-1} \cdot g_t^D$ and the dividend of asset 1 as $D_{t,1} = \xi_t \cdot \overline{D}_t$ (asset 2 pays dividend $D_{t,2} = \overline{D}_t - D_{t,1}$). The common factor g_t^D can be interpreted as a shock to the aggregate production, whereas proportions ξ_t and $1-\xi_t$ as sensitivities of different stocks (economic sectors) to this shock.

Let r_t^h denote the market share of agent h in period t by

$$r_t^h := w_t^h / \overline{w}_t$$

with $\overline{w}_t = \sum_{h \in N \cup M} w_t^h$ the aggregate wealth at t. As a direct implication of balanced growth, in any evolutionary equilibria these market shares are constant, i.e., $r_t^h \equiv r_0^h$.

Proposition 3.1 *In any steady state evolutionary equilibrium the following holds:*

1. *Asset prices p_t are proportional to the Kelly rule, i.e., for all $k \in K$*

$$\frac{p_{t,k}}{\overline{p}_t} = \mathbb{E}_t d_{t+1,k}; \tag{3.4}$$

2. *The total market capitalization $\overline{p}_t$ is given by*

$$\overline{p}_t = \overline{D}_t \cdot \beta^B / (1 - \beta^B) \tag{3.5}$$

   with a constant $\beta^B \in (0,1)$ such that $\beta^B = \beta^i \cdot \mathbb{E}_t (g_{t+1}^D)^{1-\eta^i}$ for all $i \in N$;

3. *Each broker with price impact, i.e., if*

$$(1 - f^i) r_0^i + \sum_{j \in M(i)} (1 - f^j) r_0^j \big[1 - \sum_{l \in K} \mu_l^j\big] \neq 0, \tag{3.6}$$

   invests according to the Kelly rule, i.e.,

$$\lambda^i = \mathbb{E}_t d_{t+1},$$

   otherwise his strategy is arbitrary;

[13] The assumption of stationarity can be replaced by a weaker condition of constant conditional expectations $\mathbb{E}_t (g_{t+1}^D)^{1-\eta^h}$ for each $h \in N \cup M$.

4. Each margin trader either invests according to the Kelly rule or shorts it, i.e.,

$$\mu^j = \frac{\mathbb{E}_t d_{t+1}}{\mathcal{M}} \text{ or } \mu^j = -\frac{\mathbb{E}_t d_{t+1}}{\mathcal{M}}.$$

Proof of Proposition 3.1: Follows from Theorem 3.1. □

Proposition 3.1 states that the relative market capitalization is proportional to the Kelly rule and is independent of the level of margin requirements. Margin traders follow the Kelly rule either by leveraging it long or shorting it. The broker also follows the Kelly (if his decision has price impact) and holds it non-leveraged long.[14]

Kelly prices are the only asset prices that can happen in a steady state evolutionary equilibrium. To understand this phenomenon, let us first consider the implications of the wealth of every agent growing at the same rate. We have the following result.

Proposition 3.2 *In any steady state evolutionary equilibrium:*

$$g_t^h = g_t^D \text{ and } p_{t,k}/p_{t-1,k} = g_t^D$$

for all $h \in N \cup M$, $k \in K$ and $t > 0$.

Proof of Proposition 3.2: For all $h \in N \cup M$, $g_t^h = g_t$ and $f_t^h = f^h$. Therefore (2.15) implies

$$\overline{D}_t = \sum_{h \in N \cup M} f^h w_t^h = \sum_{h \in N \cup M} f^h (g_t w_{t-1}^h) = g_t \overline{D}_{t-1},$$

which proves the first equality. Since λ_t^i and μ_t^j are constants, the market clearing condition (2.14), similarly, gives $p_{t,k} = g_t p_{t-1,k}$, completing the proof. □

The requirement of balanced growth rates entails, on the one hand, that the growth rate of the aggregate dividends is for each fund manager a benchmark net return to be delivered to clients. On the other hand, balanced growth puts some structure on the asset prices: the relative market capitalization is constant, whereas the total market capitalization increases at the rate of the aggregate dividends (which is the result of Proposition 3.2). For such asset prices, the only uncertainty an investor faces originates from unknown future dividend payments. Given that there are no redundant assets, the market portfolio is the unique portfolio that is riskless relative to the aggregate dividends, hence it is chosen by each fund manager. The Kelly prices then turn out to be the only asset prices such that the market portfolio solves both the optimization problem **[M]** of a margin trader and **[B]** of a broker.

[14] A broker has no price impact if the term in round brackets in pricing equation (2.14) is zero. This is the case for $t = 0$ when (3.6) is satisfied. In a steady state evolutionary equilibrium it also holds for subsequent $t > 0$ as market shares, fees and strategies are constant.

The argument above holds also for an evolutionary equilibrium with choice of leverage. This means that the result (3.4) on equilibrium Kelly pricing holds independently of whether or not the margin is used to the full.

That relative market capitalization is independent of margin requirements might seem unexpected at first sight. But the result immediately follows from the market clearing condition (2.14). While stricter margin requirements (higher $\mathcal{M}$) reduce the asset demand of margin traders by limiting their access to margin loans, they also increase the asset demand from brokers. Namely, when serving less margin loans, brokers are left with more funds for their own active investment. These two effects cancel each other in equilibrium. To prove this observation formally, it suffices to insert $\lambda_k^i = \mathbb{E}_t d_{t+1,k}$ and $\mu_k^j = \mathbb{E}_t d_{t+1,k}/\mathcal{M}$ into (2.14) and check that the terms containing $\mathcal{M}$ disappear.

Another interesting feature of a steady state evolutionary equilibrium is its ability to accommodate any number of long as well as short margin traders without changes to market clearing asset prices. This follows from the fact that a broker i's asset demand $x_t^i + \sum_{j\in M(i)} x_t^j$ does not depend on his clients' $j \in M(i)$ choices, as long as the clients all follow the same investment strategy as the broker (but long, leveraged-long or short). Indeed, assume that for every $i \in N$ and every $j \in M(i)$, $\mu^j = \delta^j \frac{\lambda^i}{\mathcal{M}}$ with $\delta^j = 1$ or $\delta^j = -1$. Then, using (2.3) and (2.6), we find that

$$x_{t,k}^i + \sum_{j\in M(i)} x_{t,k}^j = \left(i_t^i + \sum_{j\in M(i)} i_t^j\right)\frac{\lambda_k^i}{p_{t,k}}.$$

The market clearing condition (2.13), $\sum_{i\in N} x_{t,k}^i + \sum_{j\in M} x_{t,k}^j = 1$, implies that the market clearing asset prices

$$p_{t,k} = \sum_{i\in N}\left(i_t^i + \sum_{j\in M(i)} i_t^j\right)\lambda_k^i$$

are independent of the individual choices $\delta^j = \pm 1$ of margin traders.

3.2 Equilibrium fees and interest rates

While Proposition 3.1 provides information on asset prices and portfolio choice of agents in a steady state evolutionary equilibrium, the next Theorem 3.1 completes the characterization of these equilibria and determines the optimal operation fee rates and equilibrium interest rates.

Theorem 3.1 *The set of steady state evolutionary equilibria is uniquely characterized by the following conditions:*

1. *Investment strategies are as in Proposition 3.1;*

2. *Operation fee rates of margin traders are given by*

$$f^j = 1 - \beta^j \cdot \mathbb{E}_t(g_{t+1}^D)^{1-\eta^j} \tag{3.7}$$

and there are constants $f^S, f^L \in (0,1)$ *such that*

$$\begin{aligned} f^L &= f^j \text{ for all leveraged long traders } j \in M^L, \\ f^S &= f^j \text{ for all short sellers } j \in M^S; \end{aligned} \tag{3.8}$$

3. *Operation fee rates of brokers are given by*

$$f^i = (1-\beta^B) + \frac{1}{r_0^i} \sum_{j \in M(i)} r_0^j (1 - \beta^B - f^j) \tag{3.9}$$

with β^B *defined in Proposition 3.1;*

4. *If* $M^L \neq \emptyset$, *then*

$$1 + r = \frac{1}{1-\mathcal{M}} \frac{1}{\beta^B} - \frac{\mathcal{M}}{1-\mathcal{M}} \frac{1}{1-f^L} \tag{3.10}$$

(otherwise r *is arbitrary) and if* $M^S \neq \emptyset$, *then*

$$1 + b = \frac{1}{1+\mathcal{M}} \frac{1}{\beta^B} + \frac{\mathcal{M}}{1+\mathcal{M}} \frac{1}{1-f^S} \tag{3.11}$$

(otherwise b *is arbitrary);*

5. *For all* $i \in N$, *if* $M(i) \cap M^S \neq \emptyset$, *then*

$$r_0^i \geq \left(\frac{1-f^L}{\mathcal{M}\beta^B} \cdot \max[1-(1-\mathcal{M})\xi, 0] - 1 \right) r_0^{i,L} + \left(\frac{1-f^S}{\mathcal{M}\beta^B} - 1 \right) r_0^{i,S} \tag{3.12}$$

with $r_0^{i,L} = \sum_{j \in M(i) \cap M^L} r_0^j$ *and* $r_0^{i,S} = \sum_{j \in M(i) \cap M^S} r_0^j$.

Proof of Theorem 3.1: See Appendix B. □

Theorem 3.1 ascertains that equilibrium operation fee rates of both margin traders and brokers do not depend on the level of margin requirements $\mathcal{M}$. This finding implies that also the aggregate market capitalization is independent of margin requirements. Any change in margin requirements is absorbed by the equilibrium interest rates.

Wealth management fee rates of margin traders are determined by their discount factors, risk aversions and the growth rate of the aggregate dividends. They are identical within the class of leverage long margin traders (M^L) as well as across short sellers (M^S). An operation fee rate of a broker is determined by the balanced growth rates condition and depends on the market shares and fees of traders who have an account with this broker. Equation (3.9) implies that higher fee rates of margin traders are paired with lower operation fees of their brokers. This is a consequence of margin traders depositing their wealth under management net fees with brokers: if a trader increased an operation fee rate, he would deposit less funds with a broker. Hence the broker, who is left with less money for active investment given the same wealth under management (from his individual investors), would be forced to decrease his operation fee in order to compensate for the lost proceeds and maintain the net return at the promised rate g_t^D.

That the aggregate market capitalization is independent of margin requirements can already be conjectured from result (3.5) of Proposition 3.1. Though, since the dividends $\overline{D}_0$ are levered so that equality (2.15) holds, i.e., $\sum_{h\in N\cup M} f^h w_0^h = \overline{D}_0$, an additional step to come to this conclusion is needed. Substituting an operation fee rate f^i of a broker by its explicit expression (3.9) in (2.15), we obtain that $\overline{D}_0 = (1-\beta^B)\overline{w}_0$ with $\overline{w}_0$ the aggregate initial endowments. Consequently, the total market capitalization,

$$\overline{p}_t = \overline{D}_t \cdot \beta^B/(1-\beta^B) = \overline{D}_0 \cdot g_1^D \cdot \ldots \cdot g_t^D \cdot \beta^B/(1-\beta^B) = \overline{w}_0 \cdot g_1^D \cdot \ldots \cdot g_t^D \cdot \beta^B,$$

indeed is not affected by the exogenous margin requirements. It is remarkable that the prices of risky assets depend on time and risk preferences of brokers only. They are the same as in the case without margin trading. In fact, the equilibrium asset prices can be obtained solely from the optimization problem **[B]** of a broker, who invests into the market portfolio.[15]

Whereas margin trading has no impact on the equilibrium prices of risky assets, it affects the equilibrium interest rates. These rates are determined by the "zero profit condition" (balanced growth rates) and set competitively so that neither brokers nor margin traders can persistently deliver higher net returns to their individual clients.

Leverage gives margin traders an opportunity to amplify profits from a successful investment strategy. This advantage of traders over brokers has to be compensated by a sufficiently high interest rate on margin loans set in the market. Whether the equilibrium margin rate needs to increase or decrease in response to a higher margin requirement $\mathcal{M}$ is a priori unclear. Stricter margin requirements, on the one hand, decrease potential profits of traders by limiting the allowed leverage, and, on the other hand, reduce the total cost of their investment, as less funds are borrowed at (unchanged) margin rate. At the same time, the corresponding drop in margin loans affects brokers, who collect less interest on margin loans, but are left with more funds for their own investment.

To assess analytically the effect of the level of margin requirements on the equi-

[15] The first order condition (B.15) on the broker's investment strategy restricts the relative asset prices to be the Kelly prices, and the first order condition (B.14) on his operation fee rate determines the dividend yield of the market portfolio as a function of aggregate dividends growth rate and the broker's time and risk preferences.

librium margin rate r, we calculate the partial derivative

$$\frac{\partial r}{\partial \mathcal{M}} = \frac{1/\beta^B - 1/(1-f^L)}{(1-\mathcal{M})^2}.$$

Accordingly, since $\beta^B = \beta^i \cdot \mathbb{E}_t(g_{t+1}^D)^{1-\eta^i}$ for all $i \in N$ and $1 - f^L = \beta^j \cdot \mathbb{E}_t(g_{t+1}^D)^{1-\eta^j}$ for each $j \in M^L$, the sign of $\frac{\partial r}{\partial \mathcal{M}}$ is determined by the time and risk preferences of brokers and leveraged long margin traders. For instance, if agents have identical risk attitude, i.e., $\eta^i = \eta^j$ for all $i \in N$ and $j \in M^L$, we have $\frac{\partial r}{\partial \mathcal{M}} > 0$ if and only if $\beta^j > \beta^i$, i.e., if traders are more patient than brokers. Analogously, if agents have identical time-preferences ($\beta^i = \beta^j$ for all $i \in N$ and $j \in M$) and markets grow ($g_t^D \geq 1$), then stricter margin requirements imply higher borrowing costs ($\frac{\partial r}{\partial \mathcal{M}} > 0$) if and only if brokers are less risk-averse than margin traders, i.e., $\eta^i < \eta^j$.

Selling short, in turn, gives margin traders an opportunity to benefit from market depreciations, whereas brokers would consistently lose on their long-only investments if prices fall. This disadvantage of brokers is compensated in equilibrium by sufficiently high fees on asset loans, or equivalently, a sufficiently low rebate rate b (as it incorporates the stock loan fee).

Whether the equilibrium rebate rate increases in response to higher margin requirements, as in the case of the margin rate, is unclear. Higher margin requirements reduce short selling volume. This limits potential profits of traders in bear markets, but at the same time lowers their costs. Simultaneously, the decline of short selling volume affects negatively brokers: they collect less stock loan fees and are left with depreciating securities in hands. Analogously to the case with the margin rate, the ultimate effect of margin requirements on the equilibrium rebate rate can be examined by checking the sign of the partial derivative $\frac{\partial b}{\partial \mathcal{M}}$, which is determined by the discount factors and risk aversions of brokers and margin traders that sell short.

Explicit expressions (3.10) and (3.11) for the equilibrium interest rates reveal their dependence on agents' time and risk preferences. These characteristics have an effect on the equilibrium interest rates through their implicit impact on demand and supply of margin and asset loans. For instance, if brokers became more patient and/or less risk averse, then more margin loans and more asset loans would get available in the market. Consequently, both margin rate r and rebate rate b would decrease. Analogously, if, e.g., margin traders became more patient and/or less risk-averse, they would decrease there operation fees as to have more funds for active investments and/or would take riskier positions. The latter would raise the demand both for margin loans and deposits (recall that short sale proceeds are deposited on a margin account as collateral for the asset loan), resulting in an increased cost of borrowing r and fallen rebate rate b.

Expressions (3.10) and (3.11) also imply that the equilibrium interest are co-moving with the dividend yield of the market portfolio: when interest rates raise, the dividend yield does too, and vice versa. This follows immediately from (3.10) and (3.11) since $\frac{1}{\beta^B} = \frac{\overline{D}}{\overline{p}} + 1$ by equality (3.5).

3.2.1 Choice of business line: Whether to be a broker or a margin trader

Until now we studied evolutionary equilibria in a framework where each fund manager adheres to a certain business model, i.e., brokerage or margin trading, and makes his/her investment decisions conditional on this choice. We now examine the determinants of the choice whether to be a broker or a margin trader as well as whether to leverage or sell short for a margin trader.

Fund managers' incentives to operate in the market originate from maximizing expected utilities from consumption. In an evolutionary equilibrium the optimal consumption of an agent of any given type (a broker N, a leveraged long trader M^L or a short-seller M^S) is determined by the operation fee rate f^h with

$$c_t^h = f^h \cdot w_t^h = f^h \cdot g_1^D \cdot \ldots \cdot g_t^D \cdot w_0^h,$$

where we used that $w_t^h / w_{t-1}^h = \overline{D}_t / \overline{D}_{t-1}$ (Proposition 3.2). Accordingly, a fund manager who could choose between various business models would prefer the one with the highest operation fee rate. Consequently, for an evolutionary equilibrium to accommodate both leveraged long traders and short sellers, the condition $f^L = f^S$ has to hold. If, additionally, a broker had an option to discontinue his line of business and become a margin trader with another broker, brokers and margin traders would coexist only if $f^L = f^S = f^i$ for all $i \in N$. Summing (3.9) over $i \in N$ and rearranging the terms, this condition gives $f^L = f^S = f^i = \beta^B$. Since $\beta^B = \left(1 + \overline{D}_t/\overline{p}_t\right)^{-1}$, we observe that operation fees in this case are inverse proportional to the dividend yield of the market portfolio. The result is intuitive. Indeed, the higher the dividend yield, the more profitable the investment in risky assets, hence lower operation fees may be charged to maintain the same level of consumption.

The condition $f^L = f^S$ on coexistence of leveraged long margin traders and short sellers may be rewritten in terms of equilibrium interest rates. From (3.10) and (3.11) it follows that $f^L = f^S$ if and only if

$$b \cdot (1 + \mathcal{M}) + r \cdot (1 - \mathcal{M}) = 2 \cdot \frac{\overline{D}}{\overline{p}}.$$

Thus, when the brokerage market is perfectly competitive, i.e., $r = b = \frac{\overline{D}_t}{\overline{p}_t}$, both leveraged long margin traders and short sellers are present. Otherwise, a higher margin rate $r^\varepsilon = \frac{\overline{D}_t}{\overline{p}_t} + \varepsilon^r > \frac{\overline{D}_t}{\overline{p}_t}$ should be paired with the lower rebate rate $b^\varepsilon = \frac{\overline{D}_t}{\overline{p}_t} - \varepsilon^b < \frac{\overline{D}_t}{\overline{p}_t}$ with $\varepsilon^b = \frac{1-\mathcal{M}}{1+\mathcal{M}}\varepsilon^r$. The latter means that when borrowing funds gets more costly for a trader, depositing funds with a broker should deliver less interest too. If, however, the relation $b = b^\varepsilon$ is violated, one of the trader types disappears. When $b < b^\varepsilon$, short sellers do not receive sufficient interest on deposited short sale proceeds, hence brokers are forced to lower the operation fees as to deliver the benchmarked net return to the clients. Analogously, when $b > b^\varepsilon$ and $f^L = f^S$, selling short gives higher net returns than leveraging long. This means that short sellers can increase their operation fees without losing individual investors, hence consume more than rival leveraged long traders.

Margin rates exceed the rebate rates in real markets. This also holds in the present model. For example, when $f^L = f^S$, we have $r > b$ if and only if $\beta^j \cdot \mathbb{E}_t(g^D_{t+1})^{1-\eta^j} > \beta^i \cdot \mathbb{E}_t(g^D_{t+1})^{1-\eta^i}$ for every $j \in M$ and $i \in N$.

3.2.2 Existence of evolutionary equilibria

Theorem 3.1, which characterizes all evolutionary equilibria, also implies that exogenous parameters of the economy (e.g., discount factors, risk aversions, initial endowments) must satisfy certain conditions for an evolutionary equilibrium to exist.

Firstly, equality (3.8) implies that the discount factors and risk aversions of margin traders j, j' who are of the same type (M^L or M^S) may differ, although these two characteristics are related by the condition $\beta^j \cdot \mathbb{E}_t(g^D_{t+1})^{1-\eta^j} = \beta^{j'} \cdot \mathbb{E}_t(g^D_{t+1})^{1-\eta^{j'}}$. For instance, in a bull market ($g^D_t > 1$) this implies that $\beta^j > \beta^{j'}$ if and only if $\eta^j > \eta^{j'}$. Comparing two margin traders of the same type, we find that more patient traders are also more risk averse. The result (3.5) gives the similar condition $\beta^i \cdot \mathbb{E}_t(g^D_{t+1})^{1-\eta^i} = \beta^{i'} \cdot \mathbb{E}_t(g^D_{t+1})^{1-\eta^{i'}}$ on the discount factors and risk aversions of two brokers $i, i' \in N$.

Secondly, equation (3.9), in conjunction with inequality (3.12), restricts the set of feasible initial endowments in an evolutionary equilibrium. Applying $f^i \in (0, 1)$ to (3.9) leads to:

$$r^i_0 \geq \max\Big(\frac{1}{\beta^B} \sum_{j \in M(i)} r^j (1 - \beta^B - f^j), -\frac{1}{1-\beta^B} \sum_{j \in M(i)} r^j (1 - \beta^B - f^j)\Big),$$

giving a lower bound on the size of a broker relative to his clients. Inequality (3.12), which arises from the segregation rule (2.7), in turn, limits market shares of long margin traders and short sellers relative to their broker. Specifically, it implies that given a stricter margin requirement $\mathcal{M}$, a broker with a specified market share r^i_0 would be in the position to serve more traders of any type. This comes from the fact that stricter margin requirements, on the one hand, reduce the demand for money and asset loans from traders, and on the other hand, (by means of higher initial margins) provide the broker with more funds both for issuing margin loans and for own active investment (consequently, increasing supply of asset loans). Similarly, a higher segregation ratio ξ would allow the broker to serve more short-sellers given a fixed market share of leveraged long traders: an increase in ξ means that less margin securities need to be segregated, therefore more assets (from leveraged long margin traders) become available for short sales.

Thirdly, equality (3.7) adds further conditions to Assumption 3.1 on asset dividends. By definition $f^j \in (0,1)$, therefore (3.7) implies that the growth rate of the aggregate dividends should be such that $\beta^j \cdot \mathbb{E}_t(g^D_{t+1})^{1-\eta^j} < 1$ for all $j \in M$. For margin traders with $\eta^j \leq 1$ this holds by Assumption 3.1 (iii). For traders with $\eta^j > 1$ it would be true, if, e.g., the process g^D_t satisfied condition

$$\max_{j \in M:\, \eta^j > 1} \beta^{1/(\eta^j - 1)} < \underline{G} \leq g^D_t(s^t) \text{ for each } t > 0, s^t.$$

Finally, it is worth discussing the significance of Assumption 3.1 itself. This technical assumption, on the one hand, accounts for the existence of an evolutionary equilibrium, and, on the other hand, allows to construct it in an explicit form. In particular, time-independence of conditional relative dividends and stationarity of the growth rate of the aggregate dividends are necessary for the existence, whereas boundedness of the latter guarantees that the agents' utilities are well defined. In principle, all of these assumptions could be relaxed, and another competitive equilibrium would potentially still exist. In this general case, however, one cannot expect to find a solution in closed form. Then numerical techniques need to be employed. The numerical solution can be difficult because of the broker's optimization problem. His decisions depend on unknown and uncertain orders from the margin traders with whom he has a relationship. Conversely, if Assumption 3.1 is satisfied, then an evolutionary equilibrium that admits an explicit representation can be found analytically.

Example. The following example is an illustration of an evolutionary equilibrium. It will be further used in the simulation study on stability of evolutionary equilibria in Section 4.

Three fund managers, a broker B, a leveraged long margin trader L and a short seller S, operate in the market. The three agents have logarithmic preferences, i.e., $\eta^B = \eta^L = \eta^S = 1$, and identical discount factors $\beta^B = \beta^L = \beta^S = 99.99\%$ daily (respectively, 98% annually). The leveraged long trader and the short seller have margin accounts with the broker. The margin requirement and the segregation ratio are given by $\mathcal{M} = 50\%$ and $\xi = 140\%$ respectively.[16] The agents share an initial endowment of $\$1m$ in proportions $8:1:1$ ($w^B_0 = \$800,000$, $w^L_0 = w^S_0 = \$100,000$).

Two assets ($K = 2$) are traded daily in the market. Each period (day) there are two states ($S = 2$), which are equally likely and i.i.d. The $S \times K$ matrix of relative dividends is given by

$$d = \begin{pmatrix} 0.55 & 0.45 \\ 0.35 & 0.65 \end{pmatrix}$$

[16] These margin requirement and the segregation ratio are currently imposed by the Regulation T by the Federal Reserve and the customer protection rule (17 CFR Section 15c3-3) by the Securities and Exchange Commission in the USA.

with $\overline{D}_0 = \$195.51$ the total dividends paid at day $t = 0$ and $g_t^D = \overline{D}_{t+1}/\overline{D}_t = 100.0117\%$ the daily gross growth rate of aggregate dividends (respectively, 103% annually).[17] The margin rate and the rebate rate are set so that conditions (3.10) and (3.11) are satisfied with $r_t = (1+r) \cdot g_t^D - 1 = 0.0197\%$ and $b_t = (1+b) \cdot g_t^D - 1 = 0.0197\%$ daily (respectively, 5.10% yearly).

Given the Kelly prices (3.4)–(3.5), the investment strategies and operation fees defined in Theorem 3.1 solve the optimization problems **[B]** and **[M]** of the broker and the margin traders.[18] Conversely, given these operation fees and strategies, the Kelly prices are the market clearing prices and the growth rates of wealth are balanced. In other words, an evolutionary equilibrium is observed.

4 Stability of evolutionary equilibria: A simulation study

Evolutionary equilibrium entails a unique asset pricing forecast: relative asset prices are given by the Kelly rule. The relative market capitalization of each asset is given by the expected relative dividends, hence determined by asset fundamentals only. The outcome is rather unusual for standard asset pricing models, where equilibrium asset prices can take any positive values in response to changing characteristics of agents ("anything goes" theorem by Sonnenschein, Debreu and Mantel).[19]

Uniqueness of price forecast in evolutionary equilibria is an implication of balanced growth of wealth. The requirement that growth rates are balanced, on the one hand, prevents flows of investment money between funds (as all funds appear equally alluring to individual investors) and, on the other hand, ensures that wealth invested in a fund grows at the rate congruent with that of the average investor. For each fund manager it means that the market share under management is invariable: an investment fund formed at $t = 0$ would maintain the same market share at any future date t. This, in turn, has an implication for the market clearing asset prices: given that all agents have constant market shares, an investment decision of each particular fund manager has time-invariant degree of impact on prices.[20] Consequently, when all fund managers adhere to certain operation fee rates and investment strategies, the asset prices are stable: while absolute prices can fluctuate (due to uncertain dividend payments), the relative prices remain Kelly.

An evolutionary equilibrium however is fragile: equilibrium operation fees, investment strategies, interest rates and asset prices are related by several equality

[17] We assume that there are 252 equally spaced trading days in a year. The dividends $\overline{D}_0$ are set so that $\$50,000$ of total dividends is paid in the first year.

[18] For the operation fees we obtain: $f^L = f^S = f^B = 1 - \beta^B = 0.01\%$ daily, or equivalently, 2.05% yearly (where we accounted for wealth under management growing at $g_t^D = 103\%$ yearly). The Kelly prices (3.4)–(3.5) imply the 5.00% annual dividend yield of the market portfolio.

[19] See Mas-Colell, Whinston, and Green (1995) for the complete market case. Similar results for incomplete markets can be found in Hens and Pilgrim (2002).

[20] From equation (2.14) it follows that the bigger is the market share of an agent, the more influence his investment decision has on the prices.

conditions. This means that minor changes to its constituents may take the economy out of the evolutionary equilibrium. For instance, if one agent would deviate from the Kelly investment strategy, the market clearing prices would no longer be the Kelly prices and, as a consequence, investments into the market portfolio would no longer be optimal. While even small changes to, e.g., agents' investment strategies, can perturb the evolutionary equilibrium, it is unknown, whether this woud have a persistent effect on market shares and the Kelly pricing rule. The following simulation study aims to address this issue.

Prices and wealth dynamics without utility optimization. When running simulations, we deviate from the general equilibrium framework and focus merely on the dynamics of market clearing prices and agents' market shares given that fund managers adhere to certain wealth management fees and investment strategies.

For any specified investment strategies, operation fee rates, margin loan and rebate interest rates, the wealth dynamics (2.8)–(2.9) and the market clearing condition (2.14) generate a (non-autonomous) random dynamical system:

$$\Phi_{w_0}(t, s^{t+1}, p, w) : (p_t, w_t) \to (p_{t+1}, w_{t+1})$$

with s^{t+1} the history of states up to date $t+1$ (see Appendix C). The initial state of the system is given by endowments w_0 and market clearing asset prices p_0 determined by equation (2.14) with $i_0^h = (1 - f_0^h)w_0^h$. The mapping $(p_t, w_t) \to (p_{t+1}, w_{t+1})$ is defined well if market clearing prices p_{t+1} exist, are strictly positive and no bankruptcies happen (i.e., $w_{t+1}^h \geq 0$ for all $h \in N \cup M$). In general, this condition may be violated: e.g., a margin trader with a large market share could potentially drive the prices negative when taking a large short position in one of the assets. In the empirical study we control existence (and positivity) of prices as well occurrence of bankruptcies numerically and terminate the simulation paths where the transition from t to $t + 1$ cannot be made.[21]

Three representative agents. Though an evolutionary equilibrium can accommodate any number of agents, the corresponding asset prices may be obtained by considering only three representative agents with logarithmic utilities. Proposition 4.1, following from Theorem 3.1, states this result formally.

Proposition 4.1 *For any steady state evolutionary equilibrium, the same prices and interest rates are obtained in an steady state evolutionary equilibrium of a representative heterogeneous agent economy with the three agents:*

(i) a broker B with β^B given in Proposition 3.1, $\eta^B = 1$ and $w_0^B = \sum_{i \in N} w_0^i$,

(ii) a leveraged long trader L with $\beta^L = 1 - f^L$, $\eta^L = 1$ and $w_0^L = \sum_{j \in M^L} w_0^j$,

[21] As shown in Appendix C, existence of market clearing asset prices is equivalent to invertibility of a $K \times K$ matrix. This can be controlled by, e.g., calculating the matrix determinant. It is then straightforward to check prices positivity and presence of bankruptcies.

(iii) a short seller S with $\beta^S = 1 - f^S$, $\eta^S = 1$ and $w_0^S = \sum_{j \in M^S} w_0^j$.

Each of the three agents' portfolios are the same as the aggregate portfolios of the agents of the same type in the original economy.

4.1 Stability of Kelly pricing

We employ the example from Section 3.2.2 as a starting point of our qualitative analysis. The evolutionary equilibrium is perturbed by adding mistakes to investment strategies of the three agents and populating the market by noise traders. Mistakes in Kelly strategies and randomness of noise traders' investment decisions trigger deviations of market clearing prices from the benchmark Kelly pricing, both forces amplified by extensive leverage used. These price deviations, in turn, create opportunities for noise traders to benefit from mispricing and, when earning greater market shares, push the prices even further from fundamentals.

Formally, we model mistakes in strategy ξ by generating a vector of noise $\delta = (\delta_1, \ldots, \delta_K)$ with $\delta_k > -1$ sampled from the centered normal distribution with variance σ_m^2.[22] The variance σ_m^2 determines the size of mistakes. We then replace the strategy ξ by vector $\widetilde{\xi} = (\widetilde{\xi}_1, \ldots, \widetilde{\xi}_K)$ with $\widetilde{\xi}_k = \kappa(\xi) \cdot \xi_k(1 + \delta_k)$, where $\kappa(\xi)$ is a normalization coefficient that guarantees that the perturbed strategy has the same effective leverage and is short and long in the same assets as the original strategy ξ. To model the noise trader's strategy, we sample proportions $\mu_{t,k}$, $k \in K$ from standard normal distribution and then rescale the vector μ_t so that to achieve full usage of margin. The operation fee rate of the noise trader is set identical to other fund managers.

The simulations are conducted for two levels of margin requirements: "low" $\mathcal{M} = 40\%$ and "high" $\mathcal{M} = 60\%$. The initial market shares are set to 65%, 15%, 15% and 5% for the broker, the leveraged-long trader, the short-seller and the noise-trader respectively; agents make 5% mistakes in strategies. Table 4.1 collects the data on market characteristics calculated based on 10,000 simulations of daily trades over a 1-year horizon. The reported numbers are the average annualized trading volume, the average relative asset prices, the average absolute mispricing in relative prices (and the corresponding standard deviations), the volatility of annualized logarithmic asset returns, the average market shares, the average annualized net growth rates of wealth, the total number of market crashes and the total number of broker's bankruptcies.[23]

We find that Kelly pricing is robust to noise trading and mistakes in strategies: though prices deviate from Kelly (there is absolute mispricing in relative prices), these deviations cancel out on average. The average relative prices are the Kelly

[22] If $\delta_k \leq -1$, we resample it until $\delta_k > -1$.

[23] The average annualized net growth rate of wealth (analogue of annualized net return) is calculated as $e^{252 \cdot \overline{g}} - 1$ with $\overline{g} = \frac{1}{n} \sum_{i=1}^{n} \log(r_{t_1,i}/r_{t_0,i})/(t_1 - t_0)$ the average 1-period exponential growth rate of wealth (computed from n simulations of daily market shares between t_0 and t_1).

Table 4.1: Market characteristics when agents make 5% mistakes in strategies and noise traders are present.

Horizon		1 day		5 days		1 month		3 months		6 months		1 year	
Margin, %		40	60	40	60	40	60	40	60	40	60	40	60
Volume	Asset 1	63.64	58.41	57.87	56.44	54.08	53.83	53.04	52.56	52.69	52.19	52.30	52.02
	Asset 2	11.10	6.91	6.40	5.27	3.38	3.12	2.51	2.10	2.21	1.80	1.88	1.65
Rel. Price	Asset 1	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45
Rel. Price	Asset 1	9.03	6.81	5.60	5.40	3.51	3.80	3.06	3.12	2.81	2.94	2.58	2.71
mispr.		(0.06)	(0.04)	(0.04)	(0.04)	(0.03)	(0.03)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
	Asset 2	7.39	5.57	4.58	4.42	2.87	3.11	2.51	2.55	2.30	2.40	2.11	2.22
		(0.05)	(0.04)	(0.03)	(0.03)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
Log-Return	Asset 1	2.27	1.30	0.99	0.86	0.39	0.46	0.29	0.31	0.25	0.27	0.21	0.23
vol.	Asset 2	1.50	0.87	0.66	0.58	0.26	0.31	0.19	0.21	0.16	0.18	0.14	0.15
Mkt. Share	B	65.90	65.44	67.86	66.65	69.55	68.54	70.61	69.72	71.32	70.16	72.22	70.16
	L	15.37	15.13	16.25	15.54	17.41	16.42	18.85	17.73	20.28	19.24	21.94	21.70
	Sh	14.59	14.85	13.62	14.39	12.21	13.38	10.23	11.86	8.25	10.24	5.78	7.97
	N	4.14	4.58	2.27	3.41	0.82	1.65	0.30	0.69	0.15	0.36	0.07	0.17
Growth	B	2910.72	431.68	369.00	189.60	42.06	51.24	9.30	10.40	4.07	2.54	2.52	-0.03
	L	40004.75	795.90	1918.91	375.09	179.12	131.83	60.14	57.65	33.56	38.50	16.94	27.05
	Sh	-99.92	-92.24	-97.84	-84.67	-80.86	-67.45	-65.39	-51.43	-57.76	-44.60	-50.99	-39.46
	N	-100.00	-100.00	-100.00	-100.00	-100.00	-100.00	-99.81	-99.52	-96.07	-93.70	-85.51	-79.66
Crash	Total	0	0	30	0	35	0	35	0	35	0	35	0
	B	0	0	28	0	32	0	32	0	32	0	32	0

NOTES: The reported numbers are the average annualized trading volume in percent, the average relative asset prices, the average absolute mispricing in relative prices in percent and the corresponding standard deviations (in round brackets), the volatility of annualized logarithmic asset returns in percent, the average market shares in percent, the average annualized net growth rates of wealth, the total number of market crashes and the total number of broker's bankruptcies. The statistics are calculated based on $10,000$ simulation runs. B stands for the broker, L for the leveraged long trader, Sh for the short-seller and N for the noise trader. The market shares at $t = 0$ are given by 65%, 15%, 15% and 5% for the broker, the leveraged-long trader, the short-seller and the noise-trader respectively. The margin requirements are set to 40% and 60%. The segregation ratio is 1.4. The relative expected dividends (Kelly rule) are $(0.45, 0.55)$.

prices both in the case of "low" and "high" margin requirements, or respectively, "high" and "low" maximum leverage allowed, and over the whole observation horizon.

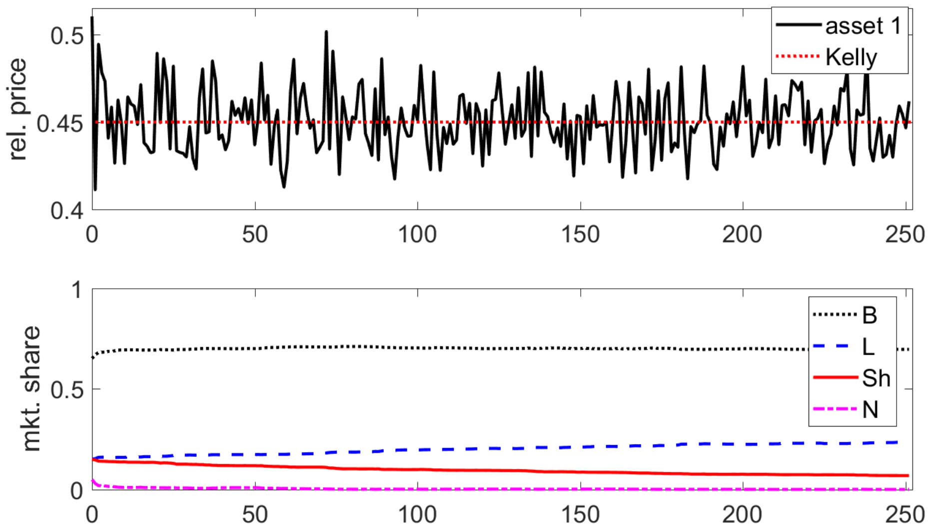
Absolute mispricing in relative prices decreases with time and is higher when more leverage is allowed. Starting with 9.03% and 6.81% after 1 trading day for $\mathcal{M} = 40\%$ and 60% respectively (asset 1), it falls more than twice to 3.51% and 3.80% after one month of trading and stabilizes to approx. 2.5% after 1 year. The degree of market mispricing is determined by the market share of the noise trader: mispricing is highest when the market share of the noise trader is largest. Long-term convergence of mispricing (to 2.58%–2.71% and 2.11%–2.22% for assets 1 and 2 respectively) is due to vanishing of noise trading, and the mispricing which remains after 1 year is caused by mistakes in strategies.[24]

Margin requirements have a clear effect on the speed of price adjustment. While relative mispricing is almost identical after one trading day (9.03% vs 6.81% for asset 1 and 7.39% vs 5.57% for asset 2), it diminishes much faster when more leverage is allowed. A similar pattern is observed in volatilities of logarithmic asset returns. The reason is that given more leverage, noise trading has more effect on the market clearing asset prices (hence we see significant mispricing and market volatility after 1 and 5 trading days), but leverage itself is destructive for noise trading. Namely, given lower margin requirements a noise trader can buy/sell more assets (trading volume is always higher when maximum leverage is higher), hence the noise trader is hit harder in case of adverse market movements. In general, the results are coherent with Friedman (1953)'s market selection hypothesis: investors trading on noise lose their money to 'smarter' traders, and thus their wealth and their impact on aggregate demand declines.

Margin requirements also affect the frequency of crashes. Over a time horizon of 1 year and 10,000 simulation runs there are no crashes when $\mathcal{M} = 60\%$, and 35 crashes when $\mathcal{M} = 40\%$ (the latter corresponds to one crash every 285 years on average). As recorded in Table 4.1, most of the crashes occur in conjunction with a broker's bankruptcy and happens within a period of 3 months. The probability of a crash is the higher, the more leverage is allowed and the more noise trading occurs on the market. This is because high leverage hastens the death of noise traders, whose losses in case of a bankruptcy are absorbed by the broker. The bankruptcy of the broker, in turn, leads to a negative asset demand from the broker, creating a downward pressure on the prices.

[24] Whether long-term mispricing (and market volatility) is higher for "low" or "high" margin requirements depends on the *relative* market shares of the agents: only mistakes in the strategies of margin traders may be amplified by leverage (the broker cannot leverage). When the market is dominated by margin traders, higher margin requirements are associated with higher market volatility and mispricing. On the other hand, the bigger is the market share of the broker, the less pronounced is the effect of margin requirements on mispricing and volatility.

Figure 4.1: Market dynamics in the presence of noise traders and mistakes.



NOTES: The plotted values are the relative price of asset 1 (relative prices of assets 1 and 2 complement to 1) and the market shares of the broker (B), the leveraged long trader (L), the short seller (Sh) and the noise trader (N). X-axis is time in days. B, L and Sh make 5% mistakes in investment strategies. The initial market shares are given by 65%, 15%, 15% and 5% for B, L, Sh and N respectively. The margin requirement is 50% and the segregation ratio is 1.4. The relative expected dividends (Kelly rule) are $(0.45, 0.55)$.
The average mispricing relative to the Kelly is 0.15% for asset 1 and -0.12% for asset 2.

4.1.1 Market dynamics

Figure 4.1 illustrates the market dynamics for a margin requirement of 50%. It shows a time series of the relative price of asset 1 (top panel) and the wealth shares of the agents (bottom panel). Coherent with the results in Table 4.1, we see that along the simulated timeseries the average relative asset price matches that implied by the Kelly rule. The average mispricing of individual assets (relative to the Kelly rule) is 0.15% and -0.12% for asset 1 and 2 respectively.

We also observe that the noise trader and the short seller consistently lose wealth relative to the broker and the leveraged long trader, with the noise trader having less than 1% of the market after only 10 days (Table 4.1 reveals similar patterns for $\mathcal{M} = 40\%$ and 60% levels). Accordingly, the impact of noise trading on asset prices diminishes with time. In the long term only mistakes in strategies contribute to mispricing: with 50% margin requirement, the 5% mistakes in strategies lead to a $(1/50\%) \cdot 5\% = 10\%$ mispricing. For asset 1 this means that the relative price is within the interval $[0.405, 0.495]$ almost all of the time – exactly what we see in Figure 4.1.

The broker and the leverage long margin trader consistently gain wealth at the expense of the short seller because asset mispricing is in favour of agents that hold the Kelly portfolio long and to the disadvantage of whose who short it. Indeed, given the market clearing prices $p_{t,k}$ with

$$(p_{t,1},\dots,p_{t,K})/\overline{p}_t = \widetilde{\lambda} \neq (\mathbb{E}_t d_{t+1,1},\dots,\mathbb{E}_t d_{t+1,K}),$$

the return from holding the Kelly portfolio long, $R^{Kelly}_{p_t}$, is on average higher than if the market prices are Kelly:

$$\begin{aligned}\mathbb{E}_t\big(R^{Kelly}_{p_t}\big) &:= \mathbb{E}_t\bigg(\sum_{k\in K}\mathbb{E}_t d_{t+1,k}\cdot\frac{D_{t+1,k}+p_{t+1,k}}{p_{t,k}}\bigg)\\ &= \bigg(\sum_{k\in K}\frac{(\mathbb{E}_t d_{t+1,k})^2}{\widetilde{\lambda}_k}+1\bigg)\cdot\mathbb{E}_t g^D_{t+1}\geq 2\mathbb{E}_t g^D_{t+1} = \mathbb{E}_t\big(R^{Kelly}_{\mathbb{E}_t d_{t+1}}\big),\end{aligned}\tag{4.1}$$

where we used Assumption 3.1 (iv) on dividends (g^D_{t+1} and d_{t+1} are uncorrelated) and Jensen's inequality.[25]

The observed advantage of the broker and the long trader relative to the short seller implies that the equilibrium margin loan and rebate rates r and b are too low for the market where the Kelly pricing holds only approximately. Recall that the interest rates in evolutionary equilibria are set so that the growth rates of wealth of the three agents are balanced. Given identical operation fees, this balance is equivalent to

$$\begin{aligned}\widetilde{R}^{Kelly}_{p_t}-(1-\mathcal{M})(1+r) &= \Big[\mathcal{M}-\frac{r^L}{r^B}(1-\mathcal{M})+\frac{r^S}{r^B}(1+\mathcal{M})\Big]\widetilde{R}^{Kelly}_{p_t}\\ &\quad+\frac{r^L}{r^B}(1+r)(1-\mathcal{M})-\frac{r^S}{r^B}(1+b)(1+\mathcal{M})\\ &= -\widetilde{R}^{Kelly}_{p_t}+(1+\mathcal{M})(1+b),\end{aligned}\tag{4.2}$$

where $\widetilde{R}^{Kelly}_{p_t} = R^{Kelly}_{p_t}\cdot\mathcal{M}/g^D_{t+1}$, the values r^L, r^B and r^S are the market shares of the leveraged long trader, the broker and the short seller, and the far-left, middle and far-right terms in (4.2) correspond to their net returns respectively. When the market prices are exactly Kelly, the two equalities in (4.2) hold and no comparative advantage is present. In turn, when the prices deviate from Kelly, the average return $\widetilde{R}^{Kelly}_{p_t}$ on the Kelly portfolio increases (equality (4.1)), leading to persistent losses of the short seller relative to the leveraged long trader and the broker. Raising both interest rates r and b would bring the market back into balance: wealth surplus generated by the leveraged long trader would be transferred to the short seller through interest payments between the broker and the margin traders.

[25] By Jensen's inequality we have $\sum_{k\in K}\frac{(\mathbb{E}_t d_{t+1,k})^2}{\widetilde{\lambda}_k}\geq 1/\big(\sum_{k\in K}\frac{\widetilde{\lambda}_k}{\mathbb{E}_t d_{t+1,k}}\cdot\mathbb{E}_t d_{t+1,k}\big) = 1$.

5 Evolutionary equilibria with choice of leverage

In Section 3 we studied evolutionary equilibria under the assumption that margin traders make full use of their margin. We now relax this assumption and look more closely into evolutionary equilibria with choice of leverage, i.e., when margin traders decide whether to leverage and, if so, to what extent.

We find that the Kelly asset pricing rule remains valid in this general case. As in the case where margins have to be used to the full, the condition of balanced growth of wealth together with the assumption of no redundant assets imply investments into the market portfolio, while CRRA utilities then limit relative market prices to Kelly (see thee discussion after Proposition 3.2). The choice of effective margin (i.e., effective leverage) has no effect on the aggregate market capitalization (absolute asset prices) either: margin m^j disappears from pricing equation (2.14) when trader j follows the same (maybe leveraged or short) investment strategy as his broker.

5.1 When is it optimal to leverage to maximum?

Whether it is optimal to leverage to maximum can be seen from the margin trader's optimization problem **[M]**. Recall that fund managers are concerned about the expected utility from consumption, hence the effective leverage is chosen so as to maximize this utility. Since both optimal investments (Kelly rule) and optimal operation fees are independent of the margin requirements (Proposition 3.1 and Theorem 3.1), the problem reduces to:

$$m^j = \arg\max_{m\in[\mathcal{M},1]} \Big\{ 1_{m\in[\mathcal{M},1)} \cdot \Big[\frac{1}{m}(1+\overline{D}/\overline{p}) - (1+r)\frac{(1-m)}{m} \Big] + 1_{m\in[1,\infty)} \cdot \Big[\frac{1}{m}(1+\overline{D}/\overline{p}) + (1+b)\frac{(m-1)}{m} \Big] \Big\} \tag{5.1}$$

for a long trader $j \in M^L$ and

$$m^j = \arg\max_{m\in[\mathcal{M},\infty)} \Big\{ -\frac{1}{m}(1+\overline{D}/\overline{p}) + (1+b)\frac{(1+m)}{m} \Big\} \tag{5.2}$$

for a short seller $j \in M^S$, where expressions in curly brackets correspond to one-period returns on Kelly investment adjusted for dividends growth g_t^D.[26]

Figure 5.1 illustrates the objective functions in (5.1) (dashed line) and (5.2) (dotted line) when the rebate rate b does not exceed the margin rate r.[27]

The results on the optimal choice of leverage by the two types of agents when $b \leq r$ are summarized in Proposition 5.1 below.

[26] As the dividend yield $\overline{D}_t/\overline{p}_t$ is constant, the index t is omitted for simplicity of presentation.

[27] In this section we restrict our attention to the only sensible case when the interest on short sale proceeds does not exceed the cost of borrowing.

Proposition 5.1 *In any steady state evolutionary equilibrium with choice of leverage, when $b \le r$ the following holds:*

1. *A long margin trader is*
   - *indifferent about size of leverage, $m \in [\mathcal{M}, \infty]$, if $b = \overline{D}/\overline{p} \le r$,*
   - *leveraged, $m \in [\mathcal{M}, 1]$, if $b < r = \overline{D}/\overline{p}$,*
   - *leveraged to the maximum, $m = \mathcal{M}$, if $\overline{D}/\overline{p} \le b \le r$ or $b \le r \le \overline{D}/\overline{p}$,*
   - *paying assets in full, $m = 1$, if $b < \overline{D}/\overline{p} < r$,*
   - *not invested in risky assets, $m = \infty$, if $\overline{D}/\overline{p} < b = r$.*
2. *A short seller is*
   - *indifferent about size of leverage, $m \in [\mathcal{M}, \infty]$, if $b = \overline{D}/\overline{p}$,*
   - *leveraged to maximum, $m = \mathcal{M}$, if $b > \overline{D}/\overline{p}$,*
   - *not invested in risky assets, $m = \infty$, if $b < \overline{D}/\overline{p}$.*

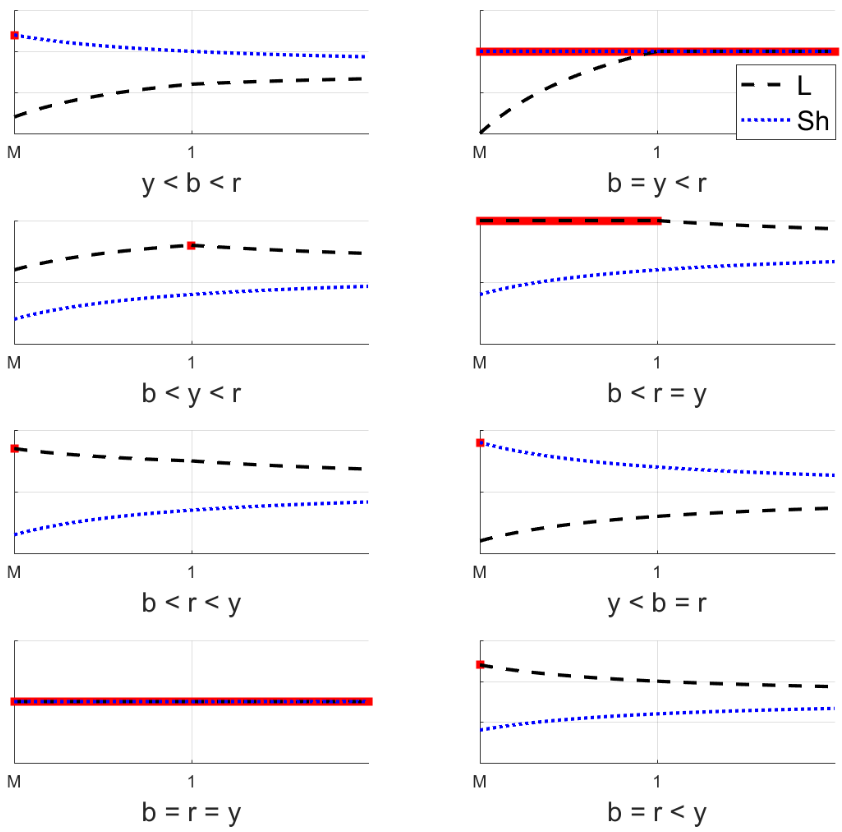
Proof of Proposition 5.1: The results follow from the first-order conditions of optimization problems (5.1) and (5.2). □

In both cases the choice of effective leverage is determined by the relation between interest rates and the dividend yield of the market portfolio. A short seller would leverage to the maximum if $b \ge \overline{D}/\overline{p}$, i.e., when the rebate rate equals or exceeds the market dividend yield. The result is very intuitive as the short seller effectively shorts the market dividend yield and profits from the interest earned on the deposited margin and the short selling proceeds. An increase in effective leverage, on the one hand, increases the short seller's adverse exposure to the market, and on the other hand gives more short sale proceeds for collecting interest on it. Accordingly, only when the interest earned on short sale proceeds is sufficiently high to compensate for losses from shorting the market portfolio, it is optimal to leverage.

In the case of a long trader, both the rebate and the margin interest rates matter for the optimal decision on how much to leverage. This is because a long trader has to pay the cost of borrowing funds r when leveraging long ($m^j \in [\mathcal{M}, 1]$) and collects the interest b on the funds that are not used in investment and deposited with a broker when $m^j \in [1, \infty)$. Accordingly, there is a tradeoff between investment into risky assets, investment into the risk-free asset (depositing funds) and usage of leverage (leverage amplifies return on investment but is subject to fees).

Formally, the optimization in (5.1) can be solved in three steps: first searching for the maximum for a nonleveraged long trader ($m \in [1, \infty)$), then for a leveraged long trader ($m \in [\mathcal{M}, 1]$) and finally choosing the solution which gives the highest objective function. Looking at the first order conditions of (5.1) for $m \in [\mathcal{M}, 1]$ one finds that a fully invested ($m \le 1$) long trader would leverage to the full if

Figure 5.1: The choice of effective margin.



NOTES: The plotted values are objective functions of a long trader (dashed line) and a short seller (dotted line) as functions of effective margins. r stands for the margin rate, b for the rebate rate and y for the dividend yield of the market portfolio. A long margin trader would leverage to maximum ($m^j = \mathcal{M}$) if $b \leq r \leq y$. A short seller would leverage to maximum if $y \leq b$. A trader who decides whether to leverage long or sell short would use margin to the full for all $b \leq r$ except $b < y < r$.

$r \leq \overline{D}/\overline{p}$, i.e., when the benefit from using leverage (expressed in collecting more returns on investment) is sufficiently high to cover the cost of borrowing. On the other hand, if $r > \overline{D}/\overline{p}$, then it is optimal not to leverage at all. The trader then needs to decide whether to make a risk-free investment instead of entering a position in risky assets, which depends on whether the rebate rate b or the market dividends yield is higher. For instance, when the rebate rate b is "low" ($b < \overline{D}/\overline{p}$), there is no advantage in a depositing funds with a broker, hence the trader would stay fully invested in risky assets, collecting the market dividend yield. If, on the contrary, the rebate rate is "high" ($b \geq \overline{D}/\overline{p}$), depositing funds with a broker becomes equally or more profitable than entering the market for risky assets. How much wealth should remain invested in risky assets then depends on the relation between r and b: e.g., when $\overline{D}/\overline{p} < b = r$, a long trader would deposit all the funds with a broker and have nothing invested in risky assets, whereas when $\overline{D}/\overline{p} < b < r$, he will spend all the funds on purchasing the market portfolio.

Comparing the maximums in optimization problems (5.1) and (5.2) of a long margin trader and a short seller, one can also derive the optimal decision of a margin trader on whether to stay (leveraged) long or short. We have the following result.

Proposition 5.2 *In any steady state evolutionary equilibrium with choice of leverage, when $b \leq r$ the following holds:*

- *if $\overline{D}/\overline{p} < b$, margin traders sell short and leverage to maximum, i.e., $m = \mathcal{M}$;*

- *if $\overline{D}/\overline{p} = b$, margin traders either sell short with $m \in [\mathcal{M}, \infty)$ or stay unlevered long with $m \in [1, \infty)$;*

- *if $b < \overline{D}/\overline{p} < r$, margin traders buy long but do not leverage, i.e., $m = 1$;*

- *if $\overline{D}/\overline{p} = r$, margin traders leverage long with $m \in [\mathcal{M}, 1]$;*

- *if $\overline{D}/\overline{p} > r$, margin traders buy long and leverage to maximum, i.e., $m = \mathcal{M}$.*

Proof of Proposition 5.2: Make use of Proposition 5.1 to compute the utility functions (5.1) and (5.2) of a long trader and a short seller respectively. Whether for given market conditions (interest rates and the market dividend yield) a margin trader prefers to short or (leverage) long is determined by which of the two utilities is higher. □

Corollary 5.1 *In an evolutionary equilibrium with choice of leverage, both leveraged long margin traders ($m < 1$) and short sellers coexist if and only if $r = b = \overline{D}/\overline{p}$.*

Proof of Corollary 5.1: Follows from Proposition 5.2. □

Figure 5.2 illustrates the results of Proposition 5.2 and Corollary 5.1. If $b = r = \overline{D}/\overline{p}$, the brokerage market is perfectly competitive and a margin trader is indifferent on whether to buy long or short and how much leverage to use. When the rebate rate is higher than the market dividend yield, i.e., $\overline{D}/\overline{p} < b \leq r$, only fully leveraged short sellers are present in the market: the rebate rate is sufficiently high to cover any potential losses from a short sale position, while leveraging long is too expensive. When the rebate rate falls below the market dividend yield, i.e., $b < \overline{D}/\overline{p}$, a long investment into the market portfolio becomes more profitable than the risk-free investment (depositing funds with a broker) as well as short selling, hence we observe long traders only, with the range of leverage depending on the ratio of r to $\overline{D}/\overline{p}$.

Figure 5.2: Decision of a margin trader whether to leverage and whether to short.

NOTES: Optimal decision of a margin trader whether to buy long or sell short and whether to leverage depending on the margin rate r, the rebate rate b and the dividend yield of the market portfolio. X-axis corresponds to different values of the market dividend yield. Y-axis is the optimal margin. The sign of the margin defines the decision to buy long (positive) or sell short (negative).

5.2 Why do margin traders both leverage long and sell short in real markets?

Corollary 5.1 implies that leveraged long traders and short sellers would coexist only in a perfectly competitive brokerage market, i.e., when $b = r = \overline{D}/\overline{p}$. This is however not the case for real markets: brokerage is not perfectly competitive, yet we still observe all types of margin traders. Within the framework of the present model, this phenomena can be explained by heterogeneity of traders' expectations regarding the (future) dividend yield of the market portfolio.

Assume that margin traders have subjective beliefs regarding the future dividend yield of the market portfolio, $\mathbb{E}_t^j(\overline{D}_{t+1}/\overline{p}_{t+1})$, but agree on the future relative market capitalization, i.e, $\mathbb{E}_t^j(p_{t+1,k}/\overline{p}_{t+1}) = \mathbb{E}_t(p_{t+1,k}/\overline{p}_{t+1})$ for all $j \in M$ and $k \in K$. Then, with minor changes in the proof of Proposition 3.1, one can show that the Kelly pricing and the Kelly investments hold in this general case too.

Subjective beliefs, though, do affect the optimal choice of leverage as well as the decision whether to short or long. This happens because margin traders effectively disagree on the future expected return of the market portfolio:[28]

$$\mathbb{E}_t^j\left(\frac{\overline{D}_{t+1} + \overline{p}_{t+1}}{\overline{p}_t}\right) = \mathbb{E}_t(g_{t+1}^D) \cdot \mathbb{E}_t^j\left(\frac{\overline{D}_t}{\overline{p}_t} + \frac{\overline{D}_t}{\overline{p}_t} / \frac{\overline{D}_{t+1}}{\overline{p}_{t+1}}\right).$$

Accordingly, when trader $j \in M$ has correct beliefs, i.e., $\mathbb{E}_t^j = \mathbb{E}_t$, his decision to short and/or leverage is determined by the relation between the market dividend yield and the two interest rates (Propositions 5.1 and 5.2). On the other hand, when the beliefs are wrong, i.e., $\mathbb{E}_t^j = \mathbb{E}_t$, these decisions are determined by the relation between

$$\widetilde{y}^j = \mathbb{E}_t^j\left(\frac{\overline{D}_t}{\overline{p}_t} + \frac{\overline{D}_t}{\overline{p}_t} / \frac{\overline{D}_{t+1}}{\overline{p}_{t+1}}\right) - 1 \tag{5.3}$$

[28] For simplicity of representation we assume that g_{t+1}^D and $\overline{D}_{t+1}/\overline{p}_{t+1}$ are independent.

and the market interest rates (Propositions 5.1 and 5.2 hold with change from $\overline{D}/\overline{p}$ to $\widetilde{y}^j$).[29]

Consequently, a margin trader j would sell short if and only if $\widetilde{y}^j \leq b$ and leverage long if and only if $b \leq \widetilde{y}^j \leq r$. This means that when $b < \overline{D}/\overline{p}$ (hence margin traders with correct beliefs would never short), short sellers believe that the current market is overpriced and trade in hope to benefit from future price declines.[30] Analogously, when $\overline{D}/\overline{p} < r$ (hence margin traders with correct beliefs would never leverage long), leveraged long traders are convinced that the market is underpriced, therefore start borrowing funds immensely to amplify the returns that they expect to achieve.

6 Conclusion and outlook

The paper introduces institutional detail of margin trading into the evolutionary finance model Evstigneev, Hens, and Schenk-Hoppé (2016, 2006, 2008, 2009, 2011). The model is used to give insights into the effects of margin requirements on (a) equilibrium asset prices and (b) market stability. We find that the Kelly pricing prediction of the evolutionary finance is robust to margin trading: margin requirements matter neither for total nor for the relative market capitalization in equilibrium. Instead there is an equilibrium relationship between margin requirements and the interest rates on borrowing and lending.

We also find that margin requirements affect the speed of price adjustment. On the one hand, when margin trading is mostly used by "smart money", the weaker the margin requirements, the quicker the correction of major mispricings and the more stable the market. On the other hand, if noise traders dominate the market, looser margin requirements lead to more market crashes and failures.

There are several aspects that are beyond the scope of this paper. Firstly, one could try to relax the assumption of balanced growth rates and replace it by, e.g., a weaker assumption of identical *expected* growth rates. This would most likely require a numerical search of an equilibrium as the extended model is rather complicated. The conjecture is that equilibrium prices might deviate from Kelly prices in the short- to medium-term but not in the long-term.

Secondly, the model does not address the issue of asset-specific margin requirements (which are common in real markets where less liquid instruments or assets with more volatile prices have higher margin requirements). In principle, it might happen that traders would decide to employ different leverage ratios across the assets. In that case investment into the market portfolio might no longer be optimal, hence the Kelly might fail in equilibrium.

[29] That a margin trader has subjective beliefs regarding the future market dividend yield induces changes in his optimization problems (5.1) and (5.2): the term $(1 + \overline{D}/\overline{p})$ gets replaced by $\widetilde{y}^j$.

[30] From (5.3) it follows that when $\widetilde{y}^j \leq b$ and $b < \overline{D}/\overline{p}$, we have $\mathbb{E}_t^j\left(\frac{\overline{D}_t}{\overline{p}_t} / \frac{\overline{D}_{t+1}}{\overline{p}_{t+1}}\right) < 1$. Similarly, when $r \leq \widetilde{y}^j$ and $\overline{D}/\overline{p} \leq r$, we have $\mathbb{E}_t^j\left(\frac{\overline{D}_t}{\overline{p}_t} / \frac{\overline{D}_{t+1}}{\overline{p}_{t+1}}\right) > 1$.

Thirdly, in real markets there is a liquidity problem associated with short selling. Having an account with a broker is not always sufficient to sell short: a trader (broker) might need to search for an asset that incurs additional costs. Moreover, illiquid assets (e.g., mortgage) can be used as collateral. In this regard, it would also be interesting to explore the relation of the present model to general equilibrium models with collateral, e.g., Brumm, Grill, Kübler, and Schmedders (2015a,b), Geanakoplos and Zame (2014).

References

Brumm, J., M. Grill, F. Kübler, and K. Schmedders (2015a). Collateral requirements and asset prices. *International Economic Review 56*, 1–25.

Brumm, J., M. Grill, F. Kübler, and K. Schmedders (2015b). Margin regulation and volatility. *Journal of Monetary Economics 73*, 54–68.

Coen-Pirani, D. (2005). Margin requirements and equilibrium asset prices. *Journal of Monetary Economics 52*, 449–475.

Evstigneev, I., T. Hens, and K. R. Schenk-Hoppé (2016). Evolutionary behavioral finance. In *The Handbook of Post Crisis Financial Modeling*, pp. 214–234. Springer.

Evstigneev, I. V., T. Hens, and K. R. Schenk-Hoppé (2006). Evolutionary stable stock markets. *Economic Theory 27*, 449–468.

Evstigneev, I. V., T. Hens, and K. R. Schenk-Hoppé (2008). Globally evolutionary stable portfolio rules. *Journal of Economic Theory 140*, 197–228.

Evstigneev, I. V., T. Hens, and K. R. Schenk-Hoppé (2009). Evolutionary finance. In T. Hens and K. R. Schenk-Hoppé (Eds.), *Handbook of Financial Markets: Dynamics and Evolution*, Chapter 9, pp. 507–566. North Holland, Amsterdam.

Evstigneev, I. V., T. Hens, and K. R. Schenk-Hoppé (2011). Local stability analysis of a stochastic evolutionary financial market model with a risk-free asset. *Mathematics and Financial Economics 5*, 185–202.

Fortune, P. (2000). Margin requirements, margin loans, and margin rates: Practice and principles. *New England Economic Review Sept/Oct*, 19–44.

Fortune, P. (2001). Margin lending and the stock market volatility. *New England Economic Review 4*, 3–26.

Friedman, M. (1953). *Essays in positive economics*. University of Chicago Press.

Geanakoplos, J. and W. R. Zame (2014). Collateral equilibrium, I: A basic framework. *Economic Theory 56*, 443–492.

Gerber, A., T. Hens, and P. Woehrmann (2010). Dynamic general equilibrium and T-period fund separation. *Journal of Financial and Quantitative Analysis 45*, 369–400.

Hakansson, N. H. (1966). *Optimal investment and consumption strategies for a class of utility functions.* Ph. D. thesis, University of California, Los Angeles.

Hakansson, N. H. (1970). Optimal investment and consumption strategies under risk for a class of utility functions. *Econometrica 38*, 587–607.

Hens, T. and B. Pilgrim (2002). *General Equilibrium Foundations of Finance: Structure of Incomplete Markets.* Kluwer Academic Publishers.

Hommes, C. (2013). *Behavioral Rationality and Heterogeneous Expectations in Complex Economic Systems.* Cambridge University Press.

Hommes, C. and F. Wagener (2009). Complex evolutionary systems in behavioral finance. In T. Hens and K. R. Schenk-Hoppé (Eds.), *Handbook of Financial Markets: Dynamics and Evolution*, Chapter 4, pp. 217–276. North Holland, Amsterdam.

Kelly, J. (1956). A new interpretation of information rate. *Bell System Technical Journal 35*, 917–926.

Long, J. (1990). The numeraire portfolio. *Journal of Financial Economics 26*, 29–69.

MacLean, L., E. Thorp, and W. T. Ziemba (Eds.) (2010). *The Kelly Capital Growth Investment Criterion: Theory and Practice*, Volume 3. World Scientific.

Mas-Colell, A., M. D. Whinston, and J. Green (1995). *Microeconomic Theory.* Oxford University Press.

Miller, M. H. and F. Modigliani (1961). Dividend policy, growth, and the valuation of shares. *Journal of Business 34*, 411–433.

Platen, E. and D. Heath (2006). *A Benchmark Approach to Quantitative Finance.* Springer.

Shiller, R. (2000). Margin calls: Should the Fed step in? *The Wall Street Journal Monday, April 10.*

Stokey, N. L. (1989). *Recursive Methods in Economic Dynamics.* Harvard University Press.

Appendix

A Segregation

The segregation rule requires that fully paid and excess margin securities are segregated. Formally, under the assumption that assets are segregated proportionally to their market value in the portfolio,

$$\Delta x^j_{t,k} = \begin{cases} x^j_{t,k} \cdot \max\left(1 + \xi \cdot \frac{1-\sum_{k\in K}\mu^j_{t,k}}{\sum_{k\in K}\max(\mu^j_{t,k},0)}, 0\right), & \text{if } x^j_{t,k} \geq 0 \text{ and } B^j_t < 0, \\ x^j_{t,k}, & \text{if } x^j_{t,k} \geq 0 \text{ and } B^j_t \geq 0, \\ 0, & \text{otherwise} \end{cases} \tag{A.1}$$

units of asset $k \in K$ must be segregated.[31]

Condition (A.1) can be deduced as follows: When the balance of the margin account is positive ($B^j_t \geq 0$), all margin securities $x^j_{t,k} \geq 0$ are considered fully paid, hence must be segregated. When the margin balance is negative ($B^j_t < 0$), only excess margin securities $\Delta x^j_{t,k} \in [0, x^j_{t,k}]$ with $x^j_{t,k} \geq 0$ and the total value of

$$\sum_{k\in K} \Delta x^j_{t,k} p_{t,k} = \max\left(\sum_{k\in K:\, x^j_{t,k}\geq 0} x^j_{t,k} p_{t,k} + \xi \cdot B^j_t, 0\right)$$

need to be segregated. Assuming that assets are segregated proportionally to their value in the portfolio, i.e., $\Delta x^j_{t,k} = \kappa^j \cdot x^j_{t,k}$, we obtain the result in (A.1).

B Evolutionary equilibria with full margin used

B.1 Restrictions on asset prices and interest rates in equilibrium

Proposition 3.2 implies that in a steady state evolutionary equilibria only asset prices with constant relative market capitalization and constant dividend yield of the market portfolio can exist, i.e, prices are of the form

$$p_{t,k} = \overline{D}_t / y^M \cdot \rho_k, \tag{B.1}$$

where $y^M > 0$ is the dividend yield and $\rho_k > 0$, $k \in K$, with $\sum_{k\in K} \rho_k = 1$, are relative prices. Further, by Lemma B.1 the real interest rates r and b are such that equations (B.4) and (B.5) admit feasible operation fee rates $f^j \in (0,1)$ of margin traders and $f^i \in (0,1)$ of brokers. When studying the optimization problems of the agents, we will assume that asset prices and interest rates satisfy these conditions.

[31] Formula (A.1) is well defined: when $B^j_t := i^j_t(1 - \sum_{k\in K}\mu^j_{t,k}) < 0$, we have $\sum_{k\in K}\max(\mu^j_{t,k}, 0) > 1$.

Lemma B.1 *In any evolutionary equilibrium the following holds:*

1. *For each margin trader $j \in M$*

$$\mu^j = \frac{1}{m^j} \cdot \frac{p_t}{\overline{p}_t} \text{ or } \mu^j = -\frac{1}{m^j} \cdot \frac{p_t}{\overline{p}_t}; \tag{B.2}$$

2. *For each broker $i \in N$ with price impact (i.e., if (3.6) holds)*

$$\lambda^i = \frac{p_t}{\overline{p}_t}, \tag{B.3}$$

otherwise it is arbitrary;

3. *For every $j \in M$ and $t \geq 0$*

$$(1+r)\cdot\min(m^j - \delta^j, 0) + (1+b)\cdot\max(m^j - \delta^j, 0) + \delta^j \cdot \left(\frac{\overline{D}_t}{\overline{p}_t} + 1\right) = \frac{m^j}{1 - f^j} \tag{B.4}$$

with $\delta^j = 1$ if $\mu^j = \frac{1}{m^j} \cdot \frac{p_t}{\overline{p}_t}$ and $\delta^j = -1$ if $\mu^j = -\frac{1}{m^j} \cdot \frac{p_t}{\overline{p}_t}$;

4. *For every $i \in N$ and $t \geq 0$*

$$r_0^i f^i + \sum_{j \in M(i)} r_0^j f^j = \left(r_0^i[1 - f^i] + \sum_{j \in M(i)} r_0^j [1 - f^j]\right) \cdot \frac{\overline{D}_t}{\overline{p}_t}. \tag{B.5}$$

Proof of Lemma B.1: 1. Applying the result $g_t^j = g_t^D$ (Proposition 3.2) to the wealth dynamics (2.10) of a margin trader $j \in M$ leads to

$$g_{t+1}^D = (1 - f^j) \cdot R_{t+1}^M(\mu^j)$$

for every $t \geq 0$. Rearranging the terms and accounting for $p_{t+1,k}/p_{t,k} = g_{t+1}^D$, $1 + r_t = g_{t+1}^D(1 + r)$ and $1 + b_t = g_{t+1}^D(1 + b)$ (Assumption 2.1), we obtain

$$\sum_{k \in K} \left\{(1 - f^j)\left[\frac{\mu_k^j}{p_{t,k}}\overline{D}_t + \overline{\mu}^j + (1+r)\cdot\min(1 - \overline{\mu}^j, 0) + (1+b)\cdot\max(1 - \overline{\mu}^j, 0)\right] - 1\right\} \frac{D_{t+1,k}}{\overline{D}_t} = 0 \tag{B.6}$$

with $\overline{\mu}^j = \sum_{l \in K} \mu_l^j$. By Assumption 3.1 (i), the $S \times K$ dimensional matrix $D_{t+1,\cdot}(s^t, \cdot)$ has rank K, therefore for each $k \in K$ the expression in curly brackets equals zero, or equivalently,

$$\frac{\mu_k^j}{p_{t,k}} \cdot \overline{D}_t = \frac{1}{(1 - f^j)} - \overline{\mu}^j - (1 + r) \cdot \min(1 - \overline{\mu}^j, 0) - (1 + b) \cdot \max(1 - \overline{\mu}^j, 0) \tag{B.7}$$

for all $k \in K$. As the right-hand side of (B.7) is independent of the index k, so must be the left-hand side. Consequently, the strategy μ^j is proportional to the relative market capitalization and satisfies (B.2).

2. Consider the dynamics of cumulative wealth of a broker $i \in N$ and his clients $j \in M(i)$. Summing (2.10) and (2.11), we have that

$$w^i_{t+1} + \sum_{j\in M(i)} w^j_{t+1} = \sum_{k\in K} \left\{ [(1-f^i)w^i_t + \overline{B^i_t}]\lambda^i_k + \sum_{j\in M(i)} (1-f^j) w^j_t \mu^j_k \right\} \frac{D_{t+1,k} + p_{t+1,k}}{p_{t,k}}$$

with $\overline{B^i_t} = \sum_{j\in M(i)} (1-f^j) w^j_t (1-\overline{\mu}^j_{t,k})$. By Proposition 3.2, $w^i_{t+1} + \sum_{j\in M(i)} w^j_{t+1} = g^D_{t+1}(w^i_t + \sum_{j\in M(i)} w^j_t)$. Inserting this into the equation above, rearranging the terms and accounting for $g^D_{t+1} = \sum_{k\in K} D_{t+1,k}/\overline{D}_t$ and $\sum_{k\in K} \lambda^i_k = 1$, we obtain that

$$\sum_{k\in K} \left\{ w^i_t \left(\frac{f^i}{\overline{D}_t} - (1-f^i)\frac{\lambda^i_k}{p_{t,k}} \right) + \sum_{j\in M(i)} w^j_t \left(\frac{f^j}{\overline{D}_t} - (1-f^j) \frac{\mu^j_k + (1-\overline{\mu}^j)\lambda^i_k}{p_{t,k}} \right) \right\} D_{t+1,k} = 0. \tag{B.8}$$

By the same argument as in part **1** this implies that $\lambda^i_k = p_{t,k}/\overline{p_t}$.

3. Follows from (B.7) by inserting the explicit expression (B.2) for μ^j_k.

4. Follows from (B.8) by inserting the explicit expressions for μ^j_k and λ^i_k and accounting for $w^h_t/\overline{w}_t \equiv r^h_0$ for all $h \in N \cup M$ and $t \geq 0$. □

B.2 Agents' optimization problems

Consider now the optimization problem **[M]** of a margin trader j and **[B]** of a broker i. For the time being we will replace the constraint (3.3) by a weaker condition

$$0 \leq f^h_t \leq 1 \tag{B.9}$$

both for brokers and margin traders (boundary cases $f^h_t = 0$ and $f^h_t = 1$ are not allowed by definition). Margin traders optimize given prices and interest rates whereas brokers take also the decisions of their clients as given.

Definition B.1 Consumption plan is *feasible* if it is nonnegative and can be achieved from the initial endowment.

Both problems have a nonempty set of feasible consumption plans. If a trader j would choose an investment strategy and an operation fee rate as in Lemma B.1 (equations (B.2) and (B.4)), the wealth under his management would grow at the rate $g^D_t > 0$ assuring nonnegative consumption $f^j w^j_t \geq 0$ in each period t. Analogously, a broker i with an operation fee rate as in (B.5) and a strategy $\lambda^i = p_t/\overline{p}_t$ would have a feasible consumption.

Lemma B.2 *For each $h \in N \cup M$ there is a constant $\overline{U}^h < \infty$, such that*

$$\mathbb{E}_0 \sum_{t=0}^{\infty} (\beta^h)^t u^h(c_t) \leq \overline{U}^h < \infty$$

there exists a constant $\overline{U}^h < \infty$*, such that on any feasible consumption plan* c*.*

Proof of Lemma B.2. By Proposition 2.1, in every period $t \geq 0$ the aggregate consumption equals the aggregate dividends $\overline{D}_t$, therefore $0 \leq c_t^h \leq \overline{D}_t$ for all $h \in N \cup M$, $t \geq 0$. Further, by Assumption 3.1 (iv), $g_t^D \leq G$. Hence, for the logarithmic utility ($\eta^h = 1$) we have:

$$\mathbb{E}_0 \sum_{t=0}^{\infty} (\beta^h)^t u^h(c_t) \leq \mathbb{E}_0 \sum_{t=0}^{\infty} (\beta^h)^t \Big[\log(\overline{D}_0) + t \log(G) \Big] \leq \frac{\mathbb{E}_0 \log(\overline{D}_0) + \beta^h \log(G)}{1 - \beta^h},$$

where we used that $\sum_{t=0}^{\infty} \beta^t = \frac{1}{1-\beta}$ and $\sum_{t=0}^{\infty} t\beta^t = \beta\big(\sum_{t=0}^{\infty} \beta^t\big)'_t = \frac{\beta}{1-\beta}$. Analogously, for a CRRA utility $u^h(c) = \frac{c^{1-\eta^h}}{1-\eta^h}$ with $\eta^h \neq 1$ we obtain:

$$\mathbb{E}_0 \sum_{t=0}^{\infty} (\beta^h)^t u^h(c_t) \leq \frac{\mathbb{E}_0(\overline{D}_0)^{1-\eta^h}}{1-\eta^h} \sum_{t=0}^{\infty} (\beta^h \mathcal{G}^{1-\eta^h})^t = \frac{\mathbb{E}_0(\overline{D}_0)^{1-\eta^h}}{1-\eta^h} \cdot \frac{1}{1 - \beta^h \mathcal{G}^{1-\eta^h}}$$

with

$$\mathcal{G} = \begin{cases} G, & \eta^h > 1, \\ \max\big[(\beta^h)^{1/(\eta^h-1)}, G\big] + 1, & \eta^h < 1, \end{cases}$$

completing the proof. □

By Lemma B.2, we have that along any feasible consumption plan c the utility function $\mathbb{E}_0 \sum_{t=0}^{\infty} (\beta^h)^t u^h(c_t)$ of agent $h \in N \cup M$ is well defined (there exists a limit of infinite sums) and bounded. Consequently, there exists a unique supremum function (the value function)

$$V_t^h(w) = \max_{f,\xi} \ \mathbb{E}_t \sum_{s=t}^{\infty} (\beta^h)^{s-t} u^h(c_s)$$

with $\xi = \mu$ for a margin trader $h \in M$ and $\xi = \lambda$ for a broker $h \in N$. By the principle of optimality[32] the supremum functions satisfy the Bellman equation **[BE]**:

$$\begin{aligned} V_t^h(w_t) = \max_{f_t, \xi_t} \ & \mathbb{E}_t \left\{ u^h(f_t w_t) + \beta^h V_{t+1}^h \left(\big[(1-f_t) w_t + 1_{h \in N} \overline{B}_t^h \big] R_{t+1}(\xi_t) + 1_{h \in N} \overline{I}_t^h \right) \right\} \\ \text{s.t.} \quad & 0 \leq f_t \leq 1, \\ & \begin{cases} \sum_{k \in K} \xi_{t,k} = 1, \ \xi_{t,k} \geq 0, \ k \in K, & \text{if } h \in N, \\ \sum_{k \in K} \xi_{t,k} = 1/\mathcal{M}, \ \xi_{t,k} \geq 0, \ k \in K, & \text{if } h \in M^L, \\ \sum_{k \in K} \xi_{t,k} = -1/\mathcal{M}, \ \xi_{t,k} \leq 0, \ k \in K, & \text{if } h \in M^S \end{cases} \end{aligned} \tag{B.10}$$

with $R_{t+1}(\xi_t) = R_{t+1}^M(\mu_t)$ if $h \in M$ and $R_{t+1}(\xi_t) = R_{t+1}^B(\lambda_t)$ if $h \in N$. These Bellman equations are necessary for an optimum. Together with the transversality

[32]No other solution can lead to an improvement of optimal policy.

condition,

$$\limsup_{t\to\infty} (\beta^h)^t V_t^h(w_t^h) \le 0, \tag{B.11}$$

the Bellman equation **[BE]** is also sufficient for an optimum.[33]

Value functions inherit concavity and differentiability from the one-period utilities $u^h(\cdot)$. Consequently, for any given $h \in N \cup M$ and $t \ge 0$ the Bellman equation **[BE]** forms a convex optimization problem with linear constraints. Therefore the Karush-Kuhn-Tucker (KKT) optimality conditions:

$$\begin{aligned} & u^{h\prime}(c_t^h) - \beta^h \cdot \mathbb{E}_t\{u^{h\prime}(c_{t+1}^h)\cdot R_{t+1}(\xi_t^h)\} + \nu_{t,0}^h - \nu_{t,1}^h = 0, \\ & 0 \le f_t^h \le 1, \\ & \nu_{t,0}^h \ge 0,\ \nu_{t,0}^h \cdot f_t^h = 0, \\ & \nu_{t,1}^h \ge 0,\ \nu_{t,1}^h \cdot (1 - f_t^h) = 0. \end{aligned} \tag{B.12}$$

and

$$\begin{aligned} & \left[(1-f_t^h)w_t^h + 1_{h\in N}\overline{B}_t^h\right] \\ & \cdot \mathbb{E}_t\left\{u^{h\prime}(c_{t+1}^h)\left[\frac{D_{t+1,k}+p_{t+1,k}}{p_{t,k}} - 1_{h\in M^L}(1+r)g_{t+1}^D - 1_{h\in M^S}(1+b)g_{t+1}^D\right]\right\} \\ & + (1_{h\in M^S} - 1_{h\in M^L\cup N})\cdot \nu_{t,m}^h + (1_{h\in M^S} - 1_{h\in M^L\cup N})\cdot \nu_{t,k}^h = 0, \\ & 1_{h\in N}\cdot \sum_{k\in K}\xi_{t,k}^h = 1_{h\in N}, \\ & (1_{h\in M^L} - 1_{h\in M^S})\cdot \sum_{k\in K}\xi_{t,k}^h = 1/\mathcal{M}, \\ & (1_{h\in M^L\cup N} - 1_{h\in M^S})\cdot \xi_{t,k}^h \le 0,\ \nu_{t,k}^h \ge 0,\ \nu_{t,k}^h\xi_{t,k}^h = 0,\ k \in K, \end{aligned} \tag{B.13}$$

are necessary and sufficient for an optimum.[34] In the above we used the envelope theorem to obtain the differential

$$\frac{\partial V_t^h}{\partial w}(w_t^h) = u^{h\prime}(c_t^h).$$

B.3 Necessary conditions for an equilibrium

By definition, $f_t^h \in (0,1)$, hence $\nu_{t,0}^h = \nu_{t,1}^h = 0$ for each $h \in N \cup M$ and $t \ge 0$. Further, $f_t^h \equiv f^h$, therefore $\frac{u^{h\prime}(c_{t+1}^h)}{u^{h\prime}(c_t^h)} = (c_{t+1}^h/c_t^h)^{-\eta^h} = (w_{t+1}^h/w_t^h)^{-\eta^h} = (g_{t+1}^D)^{-\eta^h}$, where we used the result of Proposition 3.2 for the last equality. Conditions (B.1) on

[33] See, e.g., Stokey (1989), for details.

[34] The KKT conditions are sufficient for a convex optimization problem with linear constraints. If additionally the Slater's condition is satisfied (there is a feasible point in the relative interior of the objective function), the KKT conditions become necessary for an optimum.

In our case Slater's condition is valid. For a margin trader $j \in M$ solve (B.4) with respect to f^j to obtain $f^{j*} > 0$, and a pair (f^{j*}, μ^{j*}) with μ^{j*} as in Lemma B.1 generates strictly positive consumption. For a broker $i \in N$ solve (B.5) to obtain $f^{i*} > 0$, and a pair (f^{i*}, λ^{i*}) with $\lambda_t^{i*} = p_t/\overline{p}_t$ gives strictly positive consumption (assuming that all margin traders follow (f^{j*}, μ^{j*})).

asset prices give $\frac{D_{t+1,k}+p_{t+1,k}}{p_{t,k}} = \Big(d_{t+1,k} \cdot y^M/\rho_k + 1\Big) g^D_{t+1}$ for the return on asset k with $d_{t,k} = D_{t,k}/\overline{D}_t$ the relative dividends at t. Finally, Lemma B.1 defines the investment strategies of the agents. Applying all above to the KKT conditions (B.12) and (B.13) simplifies them to:

$$\beta^h \cdot \mathbb{E}_t\big[(g^D_{t+1})^{-\eta^h} \cdot R_{t+1}(\xi^h_t)\big] = 1 \tag{B.14}$$

and

$$1_{h\notin N_0} \cdot \mathbb{E}_t\Big\{(g^D_{t+1})^{1-\eta^h}\Big(d_{t+1,k} \cdot y^M/\rho_k - 1_{h\in M^L} \cdot r - 1_{h\in M^S} \cdot b\Big)\Big\} - \tilde{\nu}^h_{t,m} = 0 \tag{B.15}$$

with N_0 the set of brokers without price impact (condition (3.6)) and a constant $\tilde{\nu}^h_t \in \mathbb{R}$.

The KKT condition (B.15) implies that $\mathbb{E}_t(d_{t+1,k}/\rho_k)$ does not depend on index k. Consequently, the asset prices are proportional to the expected relative dividends, i.e., (3.4) holds.

The KKT condition (B.14) leads to the result (3.7) on operation fees of margin traders: by Proposition 3.2, $g^D_{t+1} = g^j_{t+1} = (1 - f^j) \cdot R^M_{t+1}(\mu^j)$, or equivalently, $R^M_{t+1}(\mu^j) = g^D_{t+1}/(1 - f^j)$ for every $j \in M$. Inserting this into (B.14) gives (3.7). That $f^j = f^L$ for all $j \in M^L$ and $f^j = f^S$ for all $j \in M^S$ follows from equation (B.4) as $m^j = \mathcal{M}$ for all $j \in M$.

The KKT condition (B.14) also gives the result (3.5) on the aggregate asset prices: $R^B_{t+1}(\lambda^i) = \sum_{k\in K} \frac{D_{t+1,k}+p_{t+1,k}}{p_{t,k}}\lambda^i_k = g^D_{t+1}\sum_{k\in K}(y^M \cdot \frac{d_{t+1,k}}{\mathbb{E}_t d_{t+1,k}} + 1)\lambda^i_k$. Inserting the last expression into (B.14), using that $\sum_{k\in K}\lambda^i_k = 1$ and that there is no correlation between $(g^D_{t+1})^{1-\eta^h}$ and $d_{t+1,k}$ (by Assumption 3.1 (iv)), one obtains the equality (3.5).

Conditions (3.10) and (3.11) on interest rates follow from the clause (B.4) of Lemma B.1. The result (3.9) on operation fee rates of brokers follows from equations (B.5) and (3.5). Finally, condition (3.12) on initial endowments arises from the segregation rule (2.7).

B.4 Sufficiency of conditions for an equilibrium

Consider asset prices defined by equations (3.4)–(3.5) and interest rates as in (3.10)–(3.11). These prices are strictly positive market clearing prices (equation (2.14) is satisfied), such that wealth of each agent grows at the rate g^D_{t+1} (follows from wealth dynamics (2.10) and (2.11)), brokers have positive total asset demand (2.12) and the segregation rule works (inequality (2.7) holds). Moreover, for these prices and interest rates the given investment strategies and fees solve optimization problems **[M]** and **[B]**: the KKT conditions (B.12)–(B.13) and the transversality condition (B.11) are satisfied.[35] Consequently, the strategies, fees and interest rates determine an evolutionary equilibrium. It is in a steady state as the fees and strategies are constant.

C Market clearing prices

The t-period market clearing asset prices can be found as a solution of linear equations:

$$p_t = A_t \cdot (p_t + D_t) + v_t,$$

where A_t is a square $K \times K$ matrix with the kth row and lth column element

$$a_{t,(k,l)} = \sum_{i \in N} (1 - f_t^i) \lambda_{t,k}^i x_{t-1,l}^i + \sum_{j \in M} \varkappa_{t,k}^j x_{t-1,l}^j$$

and v_t is a K-dimensional column vector with

$$v_{t,k} = \sum_{j \in M} \big[(1+r_{t-1}) \cdot \min(B_{t-1}^j, 0) + (1+b_{t-1}) \cdot \max(B_{t-1}^j, 0)\big] \cdot [\varkappa_{t,k}^j - (1 - f_t^{i(j)}) \lambda_t^{i(j)}],$$

where $i(j) \in N$ is a broker with relation to margin trader j. The K-dimensional vector $\varkappa_k^j$ is given by:

$$\varkappa_{t,k}^j = (1 - f_t^j) \big(\mu_{t,k}^j + \big[1 - \sum_{s \in K} \mu_{t,s}^j\big] \lambda_{t,k}^{i(j)}\big).$$

The margin account balance B_{t-1}^j, the asset portfolio x_{t-1}^j of a trader and that of a broker, x_{t-1}^i, are given by expressions (2.5), (2.3) and (2.6) respectively.

[35] For the transversality condition, by Lemma B.2 we have: $\limsup_{t \to \infty} (\beta^h)^t V_t^h(w_t^h) = \limsup_{t \to \infty} \mathbb{E}_t \sum_{s=t}^{\infty} (\beta^h)^s u^h(c_s^h) = 0$.

Chapter II

Closed Form Option Pricing Under Generalized Hermite Expansions

joint work with Gabriel Drimus[a], *Erich Walter Farkas*[a,b], *and Ciprian Necula*[a,c]

Abstract. In this article we generalize the classical Edgeworth series expansion, the Gram-Charlier Type A expansion and the Gauss-Hermite expansion used in the option pricing literature. We obtain a closed-form pricing formula for European options by employing the generalized Hermite expansion for the risk-neutral density. The main advantage of the generalized expansion is that it can be applied to heavy-tailed return distributions, a case for which the standard Edgeworth expansions are not suitable. The expansion coefficients can be inferred directly from market option prices. We calibrate the model to European options on the S&P 500 index, and the results indicate that the option price data can be explained well by a risk-neutral density, whose tails are heavier than that of the normal distribution, but are not fat.

Keywords: European options, generalized Hermite series expansion, risk-neutral density, tails.
JEL classification: C63; G13.

[a]Department of Banking and Finance, University of Zürich, Switzerland
[b]Department of Mathematics, ETH Zürich, Switzerland
[c]Department of Money and Banking, DOFIN, Bucharest University of Economic Studies, Romania

1 Introduction

The risk-neutral measure, commonly referred to as the martingale measure or the state-price density is a fundamental concept in the modern financial theory. It provides an arbitrage-free method of pricing financial instruments based on a subset of market quoted prices. Semi-parametric density expansion approaches impose only structural assumptions on the functional form of the risk-neutral density (RND) and require few data for calibration, but often lack convergence for heavy-tailed return distributions. In this paper we propose a framework, which unifies a class of earlier density expansion approaches. The framework enables to derive a closed-form option pricing formula, which accounts for possibly heavy tailed return distributions and allows for efficient calibration of the RND directly to market quoted European option prices.

We introduce the generalized Hermite expansion of the RND of logarithmic returns of the underlying in the orthogonal basis of generalized Hermite polynomials around the Gaussian distribution with zero mean and finite variance. The classical Black and Scholes (1973), Merton (1973) model, the Gram-Charlier Type A expansion (Corrado and Su, 1996, 1997, Corrado, Su, et al., 1996), which is a modified version of the traditional Edgeworth series expansion (Jarrow and Rudd (1982)), and the Gauss-Hermite expansion (Necula, Drimus, and Farkas, 2016) are special cases of the generalized Hermite expansion. We extend the results in Necula, Drimus, and Farkas (2016) and derive a closed-form pricing formula for European options under the generalized Hermite expansion. Our analytical results can be applied to both traditional light-tailed return distributions as well as to heavy-tailed return distributions.[1]

The expansion coefficients can be computed from the probability density function, the characteristic function of logarithmic returns or calibrated directly to market data on prices of European options. Consequently, the method can be employed when neither the density nor characteristic function is known in closed form, such as non-affine stochastic volatility models, e.g., Kaeck and Alexander (2012), Andersen, Fusari, and Todorov (2015). Calibration of expansion coefficients to market data is conducted for a given maturity and various strike prices. Linear structure of the option pricing formula allows to use quadratic programming for calibration. The procedure can be applied to recover the whole term-structure of the RND as well as its dynamics over time.[2] The amount of data needed for calibration for one maturity (and one observation date) is determined by the number of terms (expansion coefficients) used in the expansion, usually sufficient in the range of 20–30.

[1]Heavy tails (relative to Gaussian distribution) of the RND implied by option data have been documented in, e.g., Bates (2003), Aït-Sahalia and Lo (1998), Jackwerth (2000).

[2]Since options are the most liquid derivatives and the whole cross-section of option prices is available at each observation date, they allow for a natural inference for both time and state specific preferences.

Convergence of the generalized Hermite expansion is guaranteed by a semi-parametric assumption on the tails of the RND. It requires that the standardized RND decays on its tails as $\overline{o}(x) \cdot e^{-\xi \frac{x^2}{2}}$, where $\overline{o}(x) \to 0$ when $x \to \pm\infty$. The parameter ξ may take values in $[0, \infty]$, depending on the choice of Hermite polynomials and the target distribution employed in the expansion. Thus, the Gram-Charlier Type A expansion (Corrado and Su, 1996, 1997, Corrado et al., 1996) requires that the tails of the RND decrease faster than $e^{-\frac{x^2}{4}}$ at infinity, whereas for the Gauss-Hermite expansion (Necula et al., 2016) the parameter ξ equals zero, which allows for heavy tailed as well as fat tailed distributions.

The semi-parametric generalized Hermite expansion provides a plausible alternative to both parametric and non-parametric methods to recover the RND from option prices. It imposes assumptions on the functional form of the RND's tails, but does not require explicit specification of dynamics of the underlying.[3] Truncating the infinite series after sufficiently many but finite number of expansion terms makes the generalized Hermite expansion suitable to account for stylized facts, such as non-flat "smile" and "smirk" patterns in BSM implied volatilities. Computationally efficient, the generalized Hermite expansion requires significantly less data for calibration than other model-free but data intensive non-parametric methods, such as, e.g., Breeden and Litzenberger (1978), Aït-Sahalia and Lo (1998), Aït-Sahalia and Duarte (2003).

As an important practical application, we calibrate the model to prices of S&P 500 European options in 2007-2008 and 2014–2015. We show that these prices can be explained well by the RND, whose (standardized) tails are heavier than of standard normal distribution but are not fat, i.e., decrease to zero faster than at the power law. This result is qualitatively different to findings in, e.g., Birru and Figlewski (2012) and Hamidieh (2017), who, respectively, use the Generalized Extreme Value (GEV) and the Generalized Pareto (GP) distributions for the tails of the same RND and discover presence of fat tails prior and during the fall of 2008.

Our quantitative results on the value of the tail parameter ξ are subject to the estimation methodology, but our qualitative results are significant in the context of rising interest on the tails of options implied RNDs. In particular, the asset pricing literature has recently focused on the informative content of RND's tails for future returns.[4] Evidence on the predictive power of tails of options implied RNDs

[3]The inconsistency between the Black and Scholes (1973) and Merton (1973) model (BSM) and empirical evidence encouraged development of various extensions relaxing restrictive assumptions of the BSM and considering more sophisticated dynamics of the underlying process. Examples include the jump-diffusion, pure jump models of Merton (1976), Kou (2002), Bates (1991), Madan, Carr, and Chang (1998); the stochastic-interest-rate option models of Merton (1973), Amin and Jarrow (1992); the stochastic-volatility models of Heston (1993), Hull and White (1987), Stein and Stein (1991); stochastic-volatility and stochastic-volatility jump-diffusion models of Bates (1996), Scott (1997); the models based on the Generalized Hyperbolic process or other pure-jumps Levy processes (Bibby and Sørensen, 1996) or the models with several factors with jumps both in the dynamics of the underlying and in volatility (Andersen et al., 2015).

[4]Whether the RND, hence agents' beliefs may have information content for future returns is

has been documented, among others, in Bollerslev and Todorov (2011), Vilkov and Xiao (2013), Orosi (2015), Chen, Hsieh, Huang, and Bank (2017). The common approach consists in constructing a tail loss measure using the tails of the RND recovered from market option prices by appending a certain parametric distribution, usually a traditional fat-tailed GEV or the GP distribution, to the central part of the RND. This measure is then used as a predictive variable. To this regard, our empirical study suggests that no reliable information on the tails of the RND outside of the minimum and maximum traded strike prices can be obtained from market option data: the results depend strongly on the parametric model employed in tails' estimation. A similar inference with respect to insufficiency of the GEV and GP distributions to explain physical rather than risk neutral density of returns in DJIA and Nasdaq Composite indices has been made in Malevergne, Pisarenko, and Sornette (2006).

The prototypes of the generalized Hermite expansion have long history in option pricing literature. The first application of Edgeworth series expansion to option pricing is due to Jarrow and Rudd (1982), who expanded the RND of the terminal price of the underlying in the orthogonal basis of "probabilist" Hermite polynomials around the log-normal distribution. The expansion was truncated after four terms, so that the corresponding approximate option pricing formula incorporated three adjustments to the classical Black and Scholes (1973), Merton (1973) pricing formula accounting for differences in variance, skewness and kurtosis of the underlying and the log-normal distributions. In a subsequent work Corrado and Su (1996, 1997), Corrado et al. (1996) adopted the Jarrow-Rudd framework and derived an option pricing formula using a Gram-Charlier Type A series expansion of the RND of logarithmic returns of the underlying around the Gaussian distribution. Jurczenko, Maillet, and Negréa (2004), Corrado (2007) adjusted the results of Corrado and Su (1997) to account for martingale restriction, whereas Rompolis and Tzavalis (2007) obtained a Gram-Charlier Type C expansion, an exponential form of Gram-Charlier Type A series expansion, which guaranteed positivity of the approximating probability measure.

A slight modification of the Gram-Charlier Type A series expansion was proposed by Madan and Milne (1994), who derived an expansion of the payoff function of the contingent claim for Hilbert spaces with finite bases. The approach was adopted by Abken, Madan, and Ramamurtie (1996a), Abken, Madan, and Ramamurtie (1996b) to the family of Hermite polynomials, who employed the methodology to price, respectively, Eurodollar futures options and European options on the S&P 500, with expansion truncated after four terms. A different target distribution, namely the Gamma distribution, was considered in Brenner and Eom (1997) with the expansion

itself debatable. The underlying assumptions of perfect foresight and rational expectations have been long ago questioned by financial literature, giving rise to a class of behavioural models, see, e.g., Hens and Schenk-Hoppé (2009), Hommes (2013), for overview.

in the basis of Laguerre instead of Hermite polynomials. In a latter work, Filipović, Mayerhofer, and Schneider (2013) introduced a generic framework to perform density expansions using orthonormal polynomial basis in weighted $\mathcal{L}^2$-spaces for affine models.[5] By the example of the Heston model the authors also showed a computational advantage of the Gram-Charlier Type A expansion relative to a bilateral Gamma density weight.

The rest of the paper is organized as follows. In Section 2 we present the generalized Hermite expansion of the risk-neutral density. In Section 3 we derive a pricing formula for European call options in the context of the generalized Hermite expansion and develop the calibration procedure of expansion coefficients to market quoted option prices. In Section 4 we proceed with the empirical exercise on tails of the model implied risk-neutral density. The final section summarizes the results and conclusions.

2 Density expansions based on generalized Hermite polynomials

2.1 Generalized Hermite polynomials

The generalized Hermite polynomials $H_n^{(\alpha)}(x)$ with $\alpha \in (0,1]$ are defined recursively by

$$H_{n+1}^{(\alpha)}(x) = \frac{x}{\alpha} H_n^{(\alpha)}(x) - \frac{n}{\alpha} H_{n-1}^{(\alpha)}(x) \tag{2.1}$$

with $H_0^{(\alpha)}(x) = 1$ and $H_1^{(\alpha)}(x) = \frac{x}{\alpha}$.

The polynomials (2.1) form an orthogonal basis in the weighted Hilbert space $\mathcal{L}_\alpha^2$ of measurable functions f on the real line $\mathbb{R}$ with weighting function

$$w^{(\alpha)}(x) := e^{-\frac{x^2}{2\alpha}}$$

and finite $\mathcal{L}_\alpha^2$-norm given by

$$||f||_{\mathcal{L}_\alpha^2}^2 := \int_{\mathbb{R}} |f(x)|^2 w^{(\alpha)}(x) dx < \infty$$

(see, e.g., Andrews, Askey, and Roy (1999)). Accordingly, every function $f(x) \in \mathcal{L}_\alpha^2$ can be approximated by an infinite sum

$$\sum_{n=0}^{\infty} a_n^{(\alpha)} \cdot H_n^{(\alpha)}(x)$$

[5] As opposed to the methodology of this paper, Filipović et al. (2013) restrict the weighting function of the $\mathcal{L}^2$ Hilbert space to coincide with the target distribution.

with convergence holding in $\mathcal{L}^2_\alpha$, i.e.,

$$\lim_{M\to\infty} \left|\left| f(x) - \sum_{n=0}^{M} a_n^{(\alpha)} \cdot H_n^{(\alpha)}(x) \right|\right|_{\mathcal{L}^2_\alpha} = 0.$$

The so-called "probabilists" and "physicists" Hermite polynomials, subsequently denoted by $He_n(x)$ and $H_n(x)$ respectively, are the two special cases of generalized Hermite polynomials $H_n^{(\alpha)}(x)$. For $\alpha = 1$ we recover the "probabilists" polynomials $He_n(x)$, which have been extensively used in the option pricing literature, among others, in the seminal papers of Jarrow and Rudd (1982), Corrado and Su (1996), Corrado et al. (1996) and Corrado and Su (1997). For $\alpha = 0.5$ we recover the "physicists" polynomials $H_n(x)$ employed more recently in Necula, Drimus, and Farkas (2016).

An explicit form of the generalized Hermite polynomial of the n-th order can be obtained either as the n-th derivative of the corresponding weighting function:

$$H_n^{(\alpha)}(x) = (-1)^n \cdot e^{\frac{x^2}{2\alpha}} \cdot \frac{d^n e^{-\frac{x^2}{2\alpha}}}{dx^n}, \tag{2.2}$$

or by rescaling one of the standard Hermite polynomials:

$$H_n^{(\alpha)}(x) = \alpha^{-\frac{n}{2}} He_n\left(\frac{x}{\sqrt{\alpha}}\right) = (2\alpha)^{-\frac{n}{2}} H_n\left(\frac{x}{\sqrt{2\alpha}}\right).$$

Differentiation with respect to x reduces the order of the generalized Hermite polynomial by one with

$$H_n^{(\alpha)\prime}(x) = \frac{n}{\alpha} H_{n-1}^{(\alpha)}(x). \tag{2.3}$$

The orthogonality condition implies that

$$\int_{\mathbb{R}} H_n^{(\alpha)}(x) \cdot H_m^{(\alpha)}(x) \cdot e^{-\frac{x^2}{2\alpha}} dx = \sqrt{2\pi} \cdot n! \cdot \alpha^{-n+\frac{1}{2}} \cdot \delta_{nm} \tag{2.4}$$

for every $n, m \in \mathbb{N}$, where $\delta_{nm} = 1$ if $n = m$ and $\delta_{nm} = 0$ otherwise.

2.2 Generalized Hermite expansions

Consider now an underlying asset with price S_t in period t, which pays dividends continuously. Denote by $p_\tau(x)$ the risk-neutral density (RND) of its logarithmic return $\log(S_{t+\tau}/S_t)$ over horizon τ.[6] Let μ_τ and σ_τ^2 be, respectively, the mean and the variance of the probability distribution $p_\tau(x)$.

[6]The RND of logarithmic returns $\log(S_{t+\tau}/S_t)$ might also depend on the observation date t. We will omit the index t for simplicity of notation, but all results in Sections-2.2–3.1 hold for time-dependent RNDs too. In particular, in the empirical study of Section 4 the RND is recovered for each observation date t and option maturity τ.

Assumption 2.1 *The risk-neutral density* $p_\tau(x)$ *has finite mean* $|\mu_\tau| < \infty$ *and finite variance* $\sigma_\tau^2 < \infty$.

Denote by

$$\widetilde{p}_\tau(x) := \sigma_\tau \cdot p_\tau(\mu_\tau + x \cdot \sigma_\tau)$$

the standardized density of the logarithmic return. Let $z^{(\beta)}(x)$ be the density function of centred normal distribution with finite variance $\beta > 0$:

$$z^{(\beta)}(x) := \frac{1}{\sqrt{2\pi\beta}} e^{-\frac{x^2}{2\beta}}.$$

When $\beta = 1$, the function $z^{(\beta)}(x)$ corresponds to the density function of standard normal distribution with zero mean and unit variance.

Assumption 2.2 *There exist* $\alpha \in (0, 1]$ *and* $\beta > 0$*, such that the ratio* $\frac{\widetilde{p}_\tau(x)}{z^{(\beta)}(x)}$ *has a finite* $\mathcal{L}^2_\alpha$*-norm, i.e.,*

$$\int_{\mathbb{R}} \left[\frac{\widetilde{p}_\tau(x)}{z^{(\beta)}(x)}\right]^2 \cdot w^{(\alpha)}(x)dx < \infty.$$

Assumption 2.2 implies that the ratio $\frac{\widetilde{p}_\tau(x)}{z^{(\beta)}(x)}$ can be expanded in the orthogonal basis of generalized Hermite polynomials $H_n^{(\alpha)}(x)$ with

$$\frac{\widetilde{p}_\tau(x)}{z^{(\beta)}(x)} = \sum_{n=0}^{\infty} a_{\tau,n}^{(\alpha,\beta)} \cdot H_n^{(\alpha)}(x),$$

where $\left(a_{\tau,n}^{(\alpha,\beta)}\right)_{n\in\mathbb{N}}$ is a sequence of real coefficients and convergence holds in $\mathcal{L}^2_\alpha$. The expansion

$$\widetilde{p}_\tau(x) = z^{(\beta)}(x) \cdot \sum_{n=0}^{\infty} a_{\tau,n}^{(\alpha,\beta)} \cdot H_n^{(\alpha)}(x)$$

holds in $\mathcal{L}^2_\alpha$ as well.[7] Written in terms of the original density $p_\tau(x)$, we therefore obtain:

$$p_\tau(x) = \frac{1}{\sigma_\tau} \cdot z^{(\beta)}\left(\frac{x-\mu_\tau}{\sigma_\tau}\right) \cdot \sum_{n=0}^{\infty} a_{\tau,n}^{(\alpha,\beta)} \cdot H_n^{(\alpha)}\left(\frac{x-\mu_\tau}{\sigma_\tau}\right), \tag{2.5}$$

to which we will further refer as the *generalized Hermite expansion.*

The expansion coefficients $a_{\tau,n}^{(\alpha,\beta)}$ can be computed from the probability density function $p(x)$ using the orthogonality property (2.4) of polynomials $H_n^{(\alpha)}(x)$ as

[7] Formally, since $e^{-\frac{x^2}{\beta}} \leq 1$, we have

$$||p_\tau(x) - z^{(\beta)}(x) \cdot \sum_{n=0}^{\infty} a_{\tau,n}^{(\alpha,\beta)} H_n^{(\alpha)}(x)||^2_{\mathcal{L}^2_\alpha} = \frac{1}{2\pi\beta}\int_{\mathbb{R}}\left[\frac{\widetilde{p}_\tau(x)}{z^{(\beta)}(x)} - \sum_{n=0}^{\infty} a_{\tau,n}^{(\alpha)} H_n^{(\alpha)}(x)\right]^2 e^{-\frac{x^2}{\beta}} w^{(\alpha)}(x)dx$$
$$\leq \frac{1}{2\pi\beta}\left|\left|\frac{\widetilde{p}_\tau(x)}{z^{(\beta)}(x)} - \sum_{n=0}^{\infty} a_{\tau,n}^{(\alpha,\beta)} H_n^{(\alpha)}(x)\right|\right|^2_{\mathcal{L}^2_\alpha}.$$

follows:

$$a_{\tau,n}^{(\alpha,\beta)} = \frac{\sqrt{2\pi}\alpha^{n-\frac{1}{2}}}{n!} \cdot \int_{\mathbb{R}} p_\tau(x) \cdot H_n^{(\alpha)}\left(\frac{x-\mu_\tau}{\sigma_\tau}\right) \cdot \widetilde{z}^{(\alpha,\beta)}\left(\frac{x-\mu_\tau}{\sigma_\tau}\right) dx, \tag{2.6}$$

where

$$\widetilde{z}^{(\alpha,\beta)}(x) := \sqrt{\frac{\beta}{2\pi}} e^{-\frac{x^2}{2}\left(\frac{1}{\alpha}-\frac{1}{\beta}\right)}. \tag{2.7}$$

As the risk-neutral density $p_\tau(x)$, the expansion coefficients (2.6) depend on horizon τ (and observation date t).

Necessary conditions on the risk-neutral density. The generalized Hermite expansion (2.5) requires that Assumptions 2.1 and 2.2 are satisfied. Assumption 2.1 is a standard assumption on existence of the first two moments of the RND, whereas Assumption 2.2 imposes a certain structure on the tails of the RND. As follows from Proposition 2.1 below, if the RND satisfies Assumption 2.2 and has neither extreme events in its center, nor oscillating behaviour on its tails, then the tails of the RND decrease to zero faster than $e^{\frac{x}{2}\left(\frac{1}{2\alpha}-\frac{1}{\beta}\right)}$.

Proposition 2.1 *Assume that $\widetilde{p}_\tau(x)$ satisfies Assumption 2.2 and there exists $a \in \mathbb{R}$, such that $\widetilde{p}_\tau \leq C$ on $[-a,a]$ and $\widetilde{p}_\tau(x)$ is uniformly continuous on $(-\infty,-a]\cup[a,\infty)$. Then*

$$\widetilde{p}_\tau(x) = \overline{o}(x) \cdot e^{\frac{x^2}{2}\left(\frac{1}{2\alpha}-\frac{1}{\beta}\right)}$$

with $\overline{o}(x) \to 0$ for $x \to \pm\infty$.

Proof of Proposition 2.1: The result follows immediately from Lemma A.1 in Appendix A. □

The traditional Gram-Charlier expansions employed in the literature assume that $\alpha = 1$ and $\beta = 1$, i.e., the expansion is done with the "probabilists" polynomials $He_n(x)$ and around the standard normal probability density. In this case the condition (2.5) becomes rather problematic, as it requires the standardized density $\widetilde{p}_\tau(x)$ to decrease in the tails faster than $e^{-\frac{x^2}{4}}$. This will rule out many of the heavy tailed distributions encountered in practice. In contrast, for the case of the Gauss-Hermite expansion (with $\alpha = 0.5$ and $\beta = 1$) the condition (2.5) is automatically satisfied if, e.g., the density function is bounded.[8] In general, we notice that for any given $\beta > 0$ decreasing α allows for heavier tails. Analogously, for any given $\alpha \in (0,1]$, increasing β allows for heavier tails.

2.2.1 Characteristic function representation

When using series expansions to approximate risk-neutral densities, one has to determine the martingale restriction associated with that expansion. The follow-

[8] If $\widetilde{p}(x) \leq C$ for every $x \in \mathbb{R}$, then $\int_{\mathbb{R}}[\widetilde{p}(x)]^2 dx \leq C \cdot \int_{\mathbb{R}} \widetilde{p}(x)dx = C$.

ing result on characteristic function representation allows to derive the martingale restriction for generalized Hermite expansions.

Proposition 2.2 *Consider a probability density function $p_\tau(x)$ with mean μ_τ, standard deviation σ_τ and generalized Hermite expansion coefficients $\left(a_{\tau,n}^{(\alpha,\beta)}\right)_{n\in\mathbb{N}}$ for $\beta > 0$ and $\alpha \in (\beta/2, 1]$. Then the characteristic function of $p_\tau(x)$ can be computed as*

$$\varphi_\tau(u) = e^{iu\mu_\tau - \frac{u^2\sigma_\tau^2}{2}} \cdot \sum_{n=0}^{\infty} a_{\tau,n}^{(\alpha,\beta)} \cdot i^n \cdot G_n^{(\alpha,\beta)}(u\sigma_\tau), \tag{2.8}$$

where $G_n^{(\alpha,\beta)}(x)$ is a sequence of polynomials of order n, determined by the recursive relationship:

$$G_n^{(\alpha,\beta)}(x) = x \cdot \frac{\beta}{\alpha} \cdot G_{n-1}^{(\alpha,\beta)}(x) - \frac{\beta-\alpha}{\alpha\beta} \cdot G_{n-1}^{(\alpha,\beta)\prime}(x)$$

with $G_0^{(\alpha,\beta)}(x) = 1$ and $G_1^{(\alpha,\beta)}(x) = x \cdot \frac{\beta}{\alpha}$.[9]

Proof of Proposition 3.2: See Appendix A.1.

Corollary 2.1 *In the settings of Proposition 2.2, given that r and q are, respectively, annualized continuously compounded interest rate and dividend yield, the martingale restriction associated to the generalized Hermite expansion (2.5) is:*

$$e^{\mu_\tau - (r-q)\tau + \frac{\sigma_\tau^2}{2}} \cdot \sum_{n=0}^{\infty} a_{\tau,n}^{(\alpha,\beta)} \cdot i^n \cdot G_n^{(\alpha,\beta)}(-i\sigma_\tau) = 1.$$

Proof of Corollary 2.1. The martingale condition requires $\mathbb{E}_t\left(e^{-r\tau}S_{t+\tau}\right) = e^{-q\tau}S_t$, where $\mathbb{E}_t$ is the expectation under the risk-neutral probability measure p_τ. Since $\mathbb{E}_t(e^{-r\tau}S_{t+\tau}) = e^{-r\tau} \cdot S_t \cdot \varphi_\tau(-i)$, the result follows immediately from Proposition 2.2. □

Consequently, when the expansion is truncated after first $M+1$ terms, the martingale restriction becomes a linear constraint $e^{\mu_\tau - (r-q)\tau + \frac{\sigma_\tau^2}{2}} \cdot \sum_{n=0}^{M} a_{\tau,n}^{(\alpha,\beta)} \cdot i^n \cdot G_n^{(\alpha,\beta)}(-i\sigma_\tau) = 1$ on the expansion coefficients. In a similar manner the characteristic function representation (2.8) allows to derive linear constraints, assuring that the total mass of the approximating distribution equals to one as well as that its mean and variance are equal to the mean and variance of the original risk-neutral density $p_\tau(x)$.[10]

[9] For the Gram-Charlier expansion ($\beta = \alpha = 1$) the sequence $G_n^{(\alpha)}(x)$ becomes the standard monomial sequence $1, x, x^2, \ldots$. Analogously, for the Gauss-Hermite expansion ($\beta = 1$, $\alpha = 0.5$) we have the identity $G_n^{(\alpha)}(x) = H_n^{(\alpha)}(x)$, i.e., the sequence of polynomials consists of the classical "physicists" Hermite polynomials.

[10] a) Formula (2.8) implies that when the generalized Hermite expansion (2.5) is truncated after first $M+1$ terms, the unit mass condition $\int_{\mathbb{R}} p(x)dx = \varphi_\tau(0)$ is equivalent to $\sum_{n=0}^{M} a_{\tau,n}^{(\alpha,\beta)} \cdot i^n \cdot$

3 Option pricing under generalized Hermite expansions

The generalized Hermite expansion allows to obtain a closed-form pricing formula for European options on a given underlying.[11] The pricing formula (3.1), among other cases, accounts for the classical Black-Scholes formula as well as the pricing rules inferred under the Edgeworth/Gram-Charlier expansions.

Theorem 3.1 *Assume that the log-return risk-neutral measure for time horizon τ is characterized by an annualized mean μ, an annualized standard deviation σ and generalized Hermite expansion coefficients $\left(a_{\tau,n}^{(\alpha,\beta)}\right)_{n\in\mathbb{N}}$ with $\beta > 0$ and $\alpha > \frac{\beta}{2}$.*

Then the premium at time t of a European call option with strike price K and maturity $t+\tau$ is given by

$$C\left(K,\tau,S_t,r,q,\mu,\sigma,a_n^{(\alpha,\beta)}\right) = S_t \cdot e^{-q\tau} \cdot \Pi_1^{(\alpha,\beta)} - K \cdot e^{-r\tau} \cdot \Pi_2^{(\alpha,\beta)}, \qquad (3.1)$$

where S_t is the spot price of the underlying, r is the annualized risk-free interest rate and q is the annualized continuously compounded dividend yield. The terms $\Pi_1^{(\alpha,\beta)}$ and $\Pi_2^{(\alpha,\beta)}$ are given by

$$\begin{aligned}
\Pi_1^{(\alpha,\beta)} &:= e^{\left(\mu-(r-q)+\frac{\sigma^2\beta}{2}\right)\tau} \cdot \sum_{n=0}^{\infty} a_{\tau,n}^{(\alpha,\beta)} \cdot I_n^{(\alpha,\beta)}, \\
\Pi_2^{(\alpha,\beta)} &:= \sum_{n=0}^{\infty} a_{\tau,n}^{(\alpha,\beta)} \cdot J_n^{(\alpha,\beta)},
\end{aligned}$$

where $I_n^{(\alpha,\beta)}$ and $J_n^{(\alpha,\beta)}$ are determined by the recurrence equations

$$\begin{aligned}
I_{n+1}^{(\alpha,\beta)} &= \frac{\beta}{\alpha} \cdot H_n^{(\alpha)}(-d_2) \cdot z^{(\beta)}(-d_2 - \sigma\beta\sqrt{\tau}) + \frac{\sigma\beta\sqrt{\tau}}{\alpha} \cdot I_n^{(\alpha,\beta)} + \frac{\beta-\alpha}{\alpha} \cdot \frac{n}{\alpha} \cdot I_{n-1}^{(\alpha,\beta)}, \\
J_{n+1}^{(\alpha,\beta)} &= \frac{\beta}{\alpha} \cdot H_n^{(\alpha)}(-d_2) \cdot z^{(\beta)}(-d_2) + \frac{\beta-\alpha}{\alpha} \cdot \frac{n}{\alpha} \cdot J_{n-1}^{(\alpha,\beta)}
\end{aligned}$$

with $I_0^{(\alpha,\beta)} = \mathcal{N}\left(\frac{d_2}{\sqrt{\beta}} + \sigma\sqrt{\beta\tau}\right)$, $I_1^{(\alpha,\beta)} = \frac{\beta}{\alpha} \cdot z^{(\beta)}(-d_2 - \sigma\beta\sqrt{\tau}) + \frac{\beta}{\alpha}\sigma\sqrt{\tau} \cdot \mathcal{N}\left(\frac{d_2}{\sqrt{\beta}} + \sigma\sqrt{\beta\tau}\right)$, $J_0^{(\alpha,\beta)} = \mathcal{N}\left(\frac{d_2}{\sqrt{\beta}}\right)$, $J_1^{(\alpha,\beta)} = \frac{\beta}{\alpha} \cdot z^{(\beta)}(-d_2)$, $d_1 = \frac{\log(S_t/K)+(\mu+\sigma^2)\tau}{\sigma\sqrt{\tau}}$, $d_2 = d_1 - \sigma\sqrt{\tau}$, and $\mathcal{N}(\cdot)$ the cumulative distribution function of the standard normal distribution.

Proof of Theorem 3.1: See Appendix A.2. □

Theorem 3.1 states that the price of a European call option can be found as a

$G_n^{(\alpha,\beta)}(0) = 1$. b) Given that the approximating density has unit mass (and the expansion is truncated after $M+1$ terms), the condition $\mu_\tau = -i \cdot \varphi_\tau'(0)$ on mean is equivalent to $\sum_{n=0}^{M} a_{\tau,n}^{(\alpha,\beta)} \cdot i^n \cdot G_n^{(\alpha,\beta)\prime}(0) = 0$. Analogously, given that conditions on unit mass and mean are satisfied, the condition $\sigma_\tau^2 = -\varphi_\tau''(0) + (\varphi_\tau'(0))^2$ on variance becomes $\sum_{n=0}^{M} a_{\tau,n}^{(\alpha,\beta)} \cdot i^n \cdot G_n^{(\alpha,\beta)\prime\prime}(0) = 0$.

[11] Once parameters α and β are chosen, the corresponding generalized Hermite expansion gives a unique closed-form option pricing formula. Since parameters α and β can be varied, there is, in fact, a continuum of option pricing formulas.

weighted infinite sum of generalized Hermite expansion coefficients (2.6) as

$$C\big(K,\tau,S_t,r,q,\mu,\sigma,a_n^{(\alpha,\beta)}\big) = \sum_{n=0}^{\infty} a_{\tau,n}^{(\alpha,\beta)} \cdot \big[S_t \cdot e^{\left(\mu-r+\frac{\sigma^2\beta}{2}\right)\tau} \cdot I_n^{(\alpha,\beta)} - K\cdot e^{-r\tau}\cdot J_n^{(\alpha,\beta)}\big]. \tag{3.2}$$

The corresponding price of a put option is uniquely determined by the call-put parity. To compute a cross section of option prices (3.2) on a given date, for a given maturity, but for a range of strikes, one needs to recalculate weights (in square brackets in (3.2)) but not the expansion coefficients. On the other hand, if another option maturity or observation date is considered, then new expansion coefficients must be obtained first.

A practical application of the pricing formula (3.2) requires truncation of its infinite sum after a finite number of expansion terms, which results in a so-called "truncation error". Proposition 3.1 below provides an upper bound on this error when only first $M+1$ terms are taken into account. This gives an estimate on the speed at which the truncation error decreases when more terms are included in the expansion. In particular, when, e.g., the second derivative of the ratio $\frac{\widetilde{p}_\tau(x)}{z^{(\beta)}(x)}$ exists almost everywhere and belongs to the weighted Hilbert space $\mathcal{L}^2_\alpha$ (i.e., $k=2$), the truncation error is of order $\frac{1}{\sqrt{M}}\cdot O(x)$ with $|O(x)|\leq C$ for all $x\in\mathbb{R}$ and $C\geq 0$. The latter implies that doubling the number of terms would reduce the truncation error by ca. 30%.

Proposition 3.1 *In the settings of Theorem 3.1, if there exists $k\in\mathbb{N}$, $k\geq 2$, such that $\frac{\widetilde{p}_\tau(x)}{z^{(\beta)}(x)}\in C^k(\mathbb{R})$ and $\frac{\partial^k}{\partial x^k}\frac{\widetilde{p}_\tau(x)}{z^{(\beta)}(x)}\in\mathcal{L}^2_\alpha$, then*

$$\begin{aligned}\big|C\big(K,\tau,S_t,r,q,\mu,\sigma,a_n^{(\alpha,\beta)}\big) - C_M\big(K,\tau,S_t,r,q,\mu,\sigma,a_n^{(\alpha,\beta)}\big)\big|&\\ \leq \big(S_t\cdot e^{(\mu-r)\tau}\cdot\kappa_1^{(\alpha,\beta)} + K\cdot e^{-r\tau}\cdot\kappa_2^{(\alpha,\beta)}\big)\cdot\kappa_{3,M}^{(\alpha,\beta)}&,\end{aligned}$$

where

$$C_M\big(K,\tau,S_t,r,q,\mu,\sigma,a_n^{(\alpha,\beta)}\big) = \sum_{n=0}^{M} a_{\tau,n}^{(\alpha,\beta)} \cdot \big[S_t \cdot e^{\left(\mu-r+\frac{\sigma^2\beta}{2}\right)\tau} \cdot I_n^{(\alpha,\beta)} - K\cdot e^{-r\tau}\cdot J_n^{(\alpha,\beta)}\big]$$

and

$$\begin{aligned}\kappa_1^{(\alpha,\beta)} &:= \left[N\left(d_2\sqrt{(2\alpha-\beta)/\alpha/\beta} + 2\sigma\sqrt{\tau\beta\alpha/(2\alpha-\beta)}\right)\right]^{1/2}\cdot\exp\left\{\frac{\alpha\beta\sigma^2\tau}{2\alpha-\beta}\right\},\\ \kappa_2^{(\alpha,\beta)} &:= \left[N\left(d_2\sqrt{(2\alpha-\beta)/\alpha/\beta}\right)\right]^{1/2},\\ \kappa_{3,M}^{(\alpha,\beta)} &:= \left[\frac{\alpha^{k+1/2}}{\sqrt{2\pi\beta(2\alpha-\beta)}\cdot(k-1)\cdot(M-(k-2))\cdot\ldots\cdot M}\right]^{1/2}\cdot\left|\left|\frac{\partial^k}{\partial x^k}\frac{\widetilde{p}(x)}{z^{(\beta)}(x)}\right|\right|_{\mathcal{L}^2_\alpha}.\end{aligned}$$

Proof Proposition 3.1: See Appendix A.2. □

3.1 Calibration of expansion coefficients

The option pricing formula (3.1) requires knowledge of the generalized Hermite expansion coefficients (2.6) as well as the mean and the standard deviation of the RND of logarithmic returns over a given option horizon.

If the RND, or alternatively, its characteristic function were known, the two moments and the expansion coefficients could be computed directly using formulas (2.6) and/or (2.8) respectively. In most cases, however, neither the RND nor the characteristic function are available. In the following we propose two methods to estimate the required parameters.

3.1.1 IV method

The so-called IV method infers the generalized Hermite expansion coefficients from a continuum of out-of-the-money (OTM) European call and put market option prices, $C(K, \tau, S_t)$ and $P(K, \tau, S_t)$, with given maturity τ and across strikes from zero to infinity. The moments μ_τ and σ_τ are estimated from the same option data using the method in Bakshi, Kapadia, and Madan (2003).

Proposition 3.2 *In the settings of Theorem 3.1, the generalized Hermite expansion coefficients can be computed as:*

$$
\begin{aligned}
a_{\tau,n}^{(\alpha,\beta)} =& \frac{\sqrt{2\pi}\cdot\alpha^{n-\frac{1}{2}}}{n!}\Bigg[\tilde{z}^{(\alpha,\beta)}(d_2(F_t))\cdot H_n^{(\alpha)}(-d_2(F_t))+ \\
&+e^{r\tau}\left(\int_{F_t}^{\infty} G_n^{(\alpha,\beta)}(K)\cdot C(K,\tau,S_t)dK + \int_0^{F_t} G_n^{(\alpha,\beta)}(K)\cdot P(K,\tau,S_t)dK\right)\Bigg],
\end{aligned}
\tag{3.3}
$$

where

$$
\begin{aligned}
G_n^{(\alpha,\beta)}(K) =& \frac{\tilde{z}^{(\alpha,\beta)}(d_2(K))}{K^2\sigma^2\tau}\left\{\frac{\beta-\alpha}{\alpha\beta}\left[d_2^2(K)\frac{\beta-\alpha}{\alpha\beta} - \sigma\sqrt{\tau}\cdot d_2(K) - 1\right]H_n^{(\alpha)}(-d_2(K))\right. \\
&\left. + \frac{n}{\alpha}\left[2\frac{\beta-\alpha}{\alpha\beta}d_2(K) - \sigma\sqrt{\tau}\right]H_{n-1}^{(\alpha)}(-d_2(K)) + \frac{n(n-1)}{\alpha^2}H_{n-2}^{(\alpha)}(-d_2(K))\right\},
\end{aligned}
$$

$F_t = S_t \cdot e^{(r-q)\tau}$, $d_2(K) = \frac{\log(S_t/K)+\mu\tau}{\sigma\sqrt{\tau}}$ *and* $\tilde{z}^{(\alpha,\beta)}(x)$ *is given in (2.7).*

Proof of Proposition 3.2: See Appendix Appendix A.3. □

Formula (3.3) is based on the result from Bakshi et al. (2003) that any payoff function with bounded expectation can be spanned by a continuum of OTM European calls and puts. In practice, however, only a finite number of strikes is traded, creating truncation and discretization errors to (3.3).

To circumvent the issue, one can employ curve fitting methods to implied volatilities. This approach has been intensively used in the literature, among others by Aït-Sahalia and Lo (1998), Jiang and Tian (2005) and Carr and Wu (2009). The prices of listed options are first translated into volatilities implied by the Black and Scholes model to obtain a grid of implied volatilities at different moneyness levels. A smooth function is then fitted into the grid.[12] Finally, the classical Black and Scholes formula is applied again to convert the interpolated implied volatilities to call and put option prices, which are then used in computing the integral in (3.3).[13]

3.1.2 FIT method

The so-called FIT method is an alternative to the IV method, which allows to overcome truncation and discretization errors. The expansion coefficients $(a_{\tau,n}^{(\alpha,\beta)})_{n=0}^{M}$ as well as the two (annualized) moments μ and σ are calibrated simultaneously by minimizing the total square error between model implied and market quoted option prices:

$$\min_{(a_{\tau,n}^{(\alpha,\beta)})_{n=0}^{M},\mu,\sigma} \sum_{i=1}^{N} \left| C(K_i,\tau,S_t) - C_M(K_i,\tau,S_t,r,q,\mu,\sigma,a_{\tau,n}^{(\alpha,\beta)}) \right|^2, \tag{3.4}$$

where N the number of listed option prices with time to maturity τ, K_i and $C(K_i,\tau,S_t)$ strike and call option prices of observation i, $C_M(K_i,\tau,S_t,r,q,\mu,\sigma,a_{\tau,n}^{(\alpha,\beta)})$ model implied option prices (3.2) with expansion truncated after $M+1$ terms.[14]

The linear structure of pricing formula (3.2) as well as of martingale restriction and constraints on unit mass, mean and standard deviation of the approximating RND (see discussion after Corollary 2.1) allow to employ quadratic programming with linear constraints to solve optimization problem (3.4). Formally, the FIT method can be implemented as follows:[15]

1. Select the highest order $M = 2k$ of expansion coefficients with $k \in \mathbb{N}$;[16]

2. Compute initial estimated of (annualized) mean μ and standard deviation σ from observed option prices using the Bakshi et al. (2003) approach;

[12] E.g., Jiang and Tian (2005) suggest using cubic splines for curve-fitting of implied volatilities between the maximum and the minimum available strike prices K_{max} and K_{min}: the obtained implied volatility is smooth everywhere and fits exactly the implied volatility grid. For moneyness below the lowest available moneyness level in the market, the implied volatility at the lowest strike price K_{min} is used. Analogously, for moneynes above the highest available moneyness, the implied volatility at the highest strike K_{max} is used. The implicit assumption of this method is that the RND is normal on its tails.

[13] The curve fitting procedure employs the classical Black and Scholes formula exclusively as a tool for a one-to-one mapping between option prices and implied volatilities. It makes no assumptions on validity of the Black and Scholes model for the underlying process.

[14] Put option prices should be first converted to call option prices using the call-put parity.

[15] The calibration procedure can only be performed for option maturities with at least $M+4$ observations for different strike prices: there are $(M+1)+2$ unknown parameters (expansion coefficients, the mean and the standard deviation) to be calibrated. Option maturities with insufficient price data should be excluded from analysis.

[16] Condition $M = 2k$ (M even) is necessary to avoid real roots of the approximating RND: the approximating density $\frac{1}{\sigma_\tau} \cdot z^{(\beta)}\left(\frac{x-\mu_\tau}{\sigma_\tau}\right) \cdot \sum_{n=0}^{M} a_{\tau,n}^{(\alpha,\beta)} \cdot H_n^{(\alpha)}\left(\frac{x-\mu_\tau}{\sigma_\tau}\right)$ can be positive on its both tails (i.e., when $x \to \pm\infty$) only when M is even.

3. Compute expansion coefficients $(a_n^{(\alpha,\beta)})_{n=0}^M$ by minimizing the total square error (3.4) with linear constraints (that account for the martingale restriction, unit mass and model implied mean and standard deviation);

3. Update values for μ and σ by $\widetilde{\mu} = -i \cdot \frac{\varphi_M'(0)}{\tau}$ and $\widetilde{\sigma}^2 = \frac{1}{\tau}\big(-\varphi_M''(0) + (\varphi_M'(0))^2\big)$ with $\varphi_M(0)$ defined in (2.8) and truncated after first $M+1$ terms;

4. Repeat steps $3-4$ until convergence in μ and σ achieved (or the maximum number of iterations exceeded);

5. Remove real roots if there are any: recompute expansion coefficients by solving (3.4) with (linear) constrains on positiveness of approximating density.

Remark 3.1 *The FIT method can be employed even if no initial estimates of μ and σ are known: set values of $\mu \in \mathbb{R}$ and $\sigma > 0$ arbitrary, then adjust them simultaneously to calibrating the expansion coefficients.*

Remark 3.2 *The choice of α is irrelevant for the FIT method.*

The choice of α is irrelevant for the FIT method as it effectively searches for a polynomial of a given degree M, which approximates the ratio $\frac{\widetilde{p}_\tau(x)}{z^{(\beta)}(x)}$ with the best fit to market option prices. For different values of α (but given β) the FIT method recovers expansion coefficients of the same polynomial by equivalently rewriting it in the corresponding basis of generalized Hermite polynomials $H_n^{(\alpha)}$.

4 Empirical study: Model implied tails of the RND

In the empirical study of this section we calibrate the model to market data on European options on the S&P 500 index. We find that market quoted option prices can be explained well by a (standardized) RND, whose tails converge to zero at the speed of a negative power of squared exponent (Definition 4.1 states it formally). This underlying RND can have heavier tails than the classical normal density but is never fat tailed, as opposed to the generalized extreme value and the generalized Pareto distributions recently employed in the option pricing literature to reconstruct the tails of the RND, e.g., Birru and Figlewski (2012), Vilkov and Xiao (2013).[17]

Definition 4.1 Function $f(x)$ is *of order* $e^{-\xi\frac{x^2}{2}}$ *on the tails* if

$$f(x) = \overline{o}(x) \cdot e^{-\xi\frac{x^2}{2}}, \tag{4.1}$$

[17] We recall that in the scope of this paper the RND is the risk-neutral density of logarithmic returns of the underlying, rather than the risk-neutral density of the terminal price of the underlying, as in Birru and Figlewski (2012), Vilkov and Xiao (2013). If the RND of logarithmic returns is of order $e^{-\xi\frac{x^2}{2}}$ on the tails, then the RND of the terminal price of the underlying is not fat-tailed either: it exhibits a faster than the power law decay on its tails.

where $\overline{o}(x) \to 0$ with $x \to \pm\infty$ and for every $\varepsilon > 0$, $\overline{o}(x) \cdot e^{\varepsilon x^2} \to \infty$ with $x \to +\infty$ or $x \to -\infty$. The value $\xi = \xi(f)$ is the *tail parameter* of function f.

The tail parameter $\xi(f)$ determines the negative power of squared exponent at which the heaviest tail of function f converges to zero. Accordingly, for any $\varepsilon > 0$ the function $e^{-(\xi+\varepsilon)\frac{x^2}{2}}$ decreases faster than the heaviest tail of $f(x)$, whereas $e^{-(\xi-\varepsilon)\frac{x^2}{2}}$ decays slower than both tails of $f(x)$.

For options implied RNDs (negatively skewed and leptokurtic), the tail parameter $\xi(\widetilde{p}_\tau)$ takes values in $[0, 1)$ and corresponds to the *left* tail of the RND.

4.1 Calibration of the tail parameter

Assume now that the standardized risk-neutral density $\widetilde{p}_{t,\tau}$ of logarithmic returns $\log(S_{t+\tau}/S_t)$ satisfies condition (4.1) with tail parameter $\xi(\widetilde{p}_{t,\tau}) \leq 1$. This tail parameter can be inferred from market option prices, quoted at date t and for option maturity τ as:

$$\hat{\xi}_{t,\tau} = 1 - \frac{1}{2\alpha^*}, \tag{4.2}$$

where

$$\alpha^* = \limsup\{\alpha \in [0.5, 1] \mid \lim_{M\to\infty} ||R_M^{(\alpha)}(x)||_{\mathcal{L}^2_\alpha} \text{ exists and is finite }\} \tag{4.3}$$

with approximating polynomials $R_M^{(\alpha)}(x)$ obtained by fitting model implied prices (3.2) with $\beta = 1$ (standard normal distribution as the target distribution) to quoted option prices, and the infinite sum in (3.2) truncated after $M + 1$ terms.

The intuition behind formula (4.2) is the following. For every $\varepsilon > 0$, the ratio $\frac{\widetilde{p}_{t,\tau}(x)}{e^{-\frac{x^2}{2}}}$ belongs to the weighted Hilbert space $\mathcal{L}^2_{\alpha(\varepsilon)}$ with $\alpha(\varepsilon) = \frac{1}{2(1-\xi(p_{t,\tau}))} - \varepsilon$. Therefore, the approximating polynomials,

$$R_M^{(\alpha(\varepsilon))}(x) := \sum_{n=0}^{M} a_{\tau,n}^{(\alpha(\varepsilon))} \cdot H_n^{(\alpha(\varepsilon))}(x),$$

converge to $\frac{\widetilde{p}_{t,\tau}(x)}{e^{-\frac{x^2}{2}}}$ in $\mathcal{L}^2_{\alpha(\varepsilon)}$, i.e., the limit $\lim_{M\to\infty} ||R_M^{(\alpha)}(x)||_{\mathcal{L}^2_\alpha}$ exists and is finite. On the other hand, for any $\varepsilon < 0$, the ratio $\frac{\widetilde{p}_{t,\tau}(x)}{e^{-\frac{x^2}{2}}}$ has an infinite $\mathcal{L}^2_{\alpha(\varepsilon)}$-norm, therefore the norms of approximating polynomials,

$$S_M^{(\alpha)} := \big(\sqrt{2\pi\alpha(\varepsilon)} \cdot \sum_{n=0}^{M} n! \cdot \alpha(\varepsilon)^{-n} \cdot [a_{n,\tau}^{(\alpha(\varepsilon))}]^2\big)^{1/2}, \tag{4.4}$$

are expected to diverge.[18] The Gauss-Hermite expansion ($\alpha = 0.5$), which holds for any bounded RND, ensures that the tail parameter (4.2) is well defined.

[18] We note that the expansion coefficients $(a_{\tau,n}^{(\alpha(\varepsilon))})_{n=0}^{M}$ calibrated by the FIT method are not precisely the generalized Hermite expansion coefficients (2.6): equality holds in the limit $M \to \infty$

Remark 4.1 *Convergence in (4.3) can only be measured empirically. The tail parameter (4.2) is therefore subject to the highest order M of approximating polynomials as well as the selected benchmark for convergence.*

In the following we restrict our attention to $M \leq 40$.[19] Based on the analytical results in Myller-Lebedeff (1907) and Necula et al. (2016), we employ convergence of Gauss-Hermite approximating polynomials as the benchmark for convergence of $\mathcal{L}^2_\alpha$-norms for a general case with $\alpha > 0.5$. Approximating polynomials for $\alpha > 0.5$ are obtained by first searching the approximating Gauss-Hermite polynomials of a given order and then expanding them in the corresponding basis of the generalized Hermite polynomials $H_n^{(\alpha)}(x)$.

4.2 Data description

The study is based on the Chicago Board Options Exchange (CBOE) market for S&P 500 (SPX) index options. Options written on this index are the most actively traded European style contracts and have been in the focus of many existing investigations.

SPX options expire on the third Friday of the expiration month, with expiration months up to 12 near-term months. The multiplier is \$100, the strike price intervals are 5 points and 25 points for far months. SPX are AM settled options, trading in SPX will ordinarily cease on the business day (usually a Thursday) preceding the exercise date.[20]

The sample period is from January 1st, 2014 to December 31st, 2015. SPX options for this period are liquid, and tail parameters for both short and long option maturities can be retrieved.[21] The daily data for option prices (best bid and best ask), the trading volume, the spot price (closing), the dividend yield and the term structure of interest rates are collected from Option Metrics. The interest rates for corresponding option maturities are obtained by linear interpolation between the two closest (in maturity) zero-coupon rates on the zero curve for a given observation date.[22] If the zero curve for a given observation date is not available, then the zero curve for the closest previous observation date is used.

only. Alternatively, one could construct approximating polynomials using the generalized Hermite expansion coefficients computed by formula (3.2) (IV method). In the latter case, however, the expansion coefficients, hence the norms of approximating polynomials, would be consistently under or overestimated due to truncation and discretization errors in (3.2), see discussion after Proposition 3.2.

[19] The calibration procedure is run in MATLAB. We noticed that inclusion of more expansion terms makes the results on high order expansion coefficients unstable, which stems from the limited MATLAB precision.

[20] AM settlement means stop trading on Thursday evening, but the settlement price is not determined until Friday AM.

[21] In Appendix B we also provide results for the sample period from January 1st, 2007 to December 31st, 2008. The results are qualitatively the same, but the tail parameter for long option maturities could not be recovered.

[22] The Actual/365 day-count convention is used both for zero coupon and dividend yields.

Table 4.1: Sample properties of S&P 500 index options in 2014-2015.

		Moneyness S/K	Days-to-expiration			Subtotal
			≤ 60	$(60, 180]$	> 180	
Calls	OTM	≤ 0.94	0.58	2.41	27.42	
			(0.57)	(0.90)	(3.67)	
			[296.10]	[159.54]	[47.06]	[109.69]
			{11084}	{21668}	{50235}	{82987}
		$(0.94, 0.97]$	2.76	14.11	82.41	
			(0.73)	(1.47)	(5.13)	
			[1049.46]	[430.38]	[90.67]	[571.12]
			{11466}	{10222}	{8421}	{30109}
	ATM	$(0.97, 1.00]$	14.24	37.06	114.46	
			(1.26)	(1.94)	(5.77)	
			[2414.80]	[493.66]	[127.43]	[1193.61]
			{12334}	{9831}	{7673}	{29838}
		$(1.00, 1.03]$	47.38	72.09	146.33	
			(2.27)	(2.38)	(6.09)	
			[1397.23]	[774.50]	[139.15]	[860.02]
			{11717}	{9507}	{7604}	{28828}
	ITM	$(1.03, 1.06]$	93.99	112.95	181.06	
			(3.03)	(2.80)	(6.48)	
			[71.81]	[32.99]	[28.42]	[47.31]
			{10526}	{8723}	{7036}	{26285}
		> 1.06	451.73	478.78	679.02	
			(3.45)	(3.60)	(7.43)	
			[5.59]	[4.62]	[2.90]	[4.15]
			{87170}	{88187}	{134008}	{309365}
Puts	ITM	≤ 0.94	252.65	272.69	422.79	
			(3.81)	(3.92)	(7.91)	
			[16.38]	[12.73]	[1.90]	[7.12]
			{12983}	{18347}	{42792}	{74122}
		$(0.94, 0.97]$	99.43	117.19	208.09	
			(3.35)	(3.25)	(6.58)	
			[61.83]	[16.50]	[10.55]	[31.06]
			{10441}	{10210}	{8421}	{29072}
	ATM	$(0.97, 1.00]$	47.85	76.50	176.65	
			(2.57)	(2.57)	(6.46)	
			[765.66]	[259.16]	[74.59]	[419.48]
			{12197}	{9831}	{7673}	{29701}
		$(1.00, 1.03]$	21.12	51.66	149.13	
			(1.51)	(2.18)	(6.11)	
			[3094.02]	[1016.05]	[201.45]	[1653.91]
			{11880}	{9507}	{7604}	{28991}
	OTM	$(1.03, 1.06]$	11.09	35.91	127.81	
			(1.22)	(1.99)	(5.86)	
			[1919.26]	[452.80]	[136.40]	[973.28]
			{11024}	{8723}	{7036}	{26783}
		> 1.06	2.00	6.38	31.79	
			(0.66)	(1.03)	(3.50)	
			[457.38]	[179.56]	[50.92]	[201.46]
			{100142}	{110256}	{154205}	{364603}
Subtotal			{302964}	{315012}	{442708}	{1060684}

NOTES: The reported numbers are the average $ quoted bid-ask mid point prices, the average $ bid-ask spreads (best bid minus best ask price) in parenthesis, the average trading volumes per contract per day in square brackets and the total number of observations in curly brackets. The sample period is from January 1st, 2014 to December 31st, 2015 with total of $1,060,684$ options. S denotes the spot S&P 500 index level and K is the strike price. OTM, ATM and ITM stand for out-of-the-money, at-the-money and in-the-money options respectively.

The beginning data sample contains $1,282,632$ observations of call and put option prices. Option prices are taken as midpoints of best bid-ask quotations to avoid bid-ask bounce problems in transaction prices. Observations with mid prices less than $1/8$ are dropped and the data is filtered so that basic arbitrage constraints,

$$C(K, \tau, S_t) > \max(S_t \cdot e^{-q\tau} - K \cdot e^{-r\tau}, 0),$$
$$P(K, \tau, S_t) > \max(K \cdot e^{-r\tau} - S_t \cdot e^{-q\tau}, 0),$$

are observed. Based on these criteria, we eliminate $221,948$ observations (mostly deep in-the-money puts and calls) and this is the starting point of our analysis.

Table 4.1 describes the distribution of sampled option data along moneyness and maturities. A call (put) option is said to be out-of-the-money (OTM) if $S_t/K \leq 0.97$ ($S_t/K > 1.03$), at-the-money (ATM) if $S_t/K \in (0.97, 1.03]$ and in-the-money if $S_t/K > 1.03$ ($S_t/K \leq 0.94$). Applying a finer partition, the data is further subdivided into six moneyness categories. By time-to-expiration we classify the options as short-term (≤ 60 days to expiration), medium-term ($(60 - 180]$ days) and long-term (> 180 days). For each moneyness-maturity category the summary statistics are reported for the average mid prices, the average bid-ask spread (best bid minus best ask price), the average trading volume and the total number of observations.

There are in total $1,060,684$ observations, with 31.64% (9.73%) ITM call (put) options, 5.53% (5.53%) ATM calls (puts) and 10.66% (36.90%) OTM call (put) option observations. The average option prices vary from \$0.58 for short-term deep OTM call options to \$679.02 for long-term deep ITM calls, with the average bid-ask spread deviating from \$0.57 (for short-term deep OTM call options) to \$7.91 (for long-term deep ITM puts). The average trading volume is not persistent across the moneyness-maturity classes either, starting with 1.90 contracts per day for long-term deep ITM put options and surging to 3094.02 for short-term ATM puts.

For each moneyness class we observe that ITM options are very infrequently traded relative to ATM and OTM options. Thus the average trading volume (across all maturities) of ITM calls (puts) is 7.53 (13.86) contracts per day in contrast to ATM and OTM call (put) options with the average volume of 504.81 (355.33). This reflects a strong demand for protective puts and calls, as well as signals that ITM option prices are notoriously unreliable.

To account for this issue, we follow the approach in Aït-Sahalia and Lo (1998) and replace the prices of all illiquid options by prices implied by the put-call parity at the relevant strikes and maturities. Specifically, for each moneyness-maturity category we first discard the option data of the less liquid option type in the category (e.g., for $S_t/K \in (0.97, 1.00]$ and > 180 days to expiration the put option data is discarded). The call option prices which are illiquid are reconstructed using liquid put option prices $P(K, \tau, S_t)$ with the corresponding strike K and maturity τ using the call-put

Table 4.2: **MSE of the Gauss-Hermite expansion on June 19th, 2015.**

Days-to-exp.\ M	10	12	14	16	18	20	22	24	26	28	30	32	34	36	38	40
≤ 60	3.82	2.69	2.44	2.12	1.89	1.80	1.54	1.40	1.29	1.20	1.11	1.11	1.01	0.95	0.92	0.93
$(60, 180]$	7.80	4.47	3.47	2.22	1.98	1.37	1.27	0.98	0.92	0.76	0.70	0.62	0.57	0.54	0.48	0.48
> 180	44.81	15.73	8.85	2.02	6.71	2.74	3.41	1.76	1.90	1.46	1.26	0.99	0.87	0.72	0.65	0.57
all	50.91	18.06	10.28	2.51	7.73	3.23	3.97	2.09	2.24	1.73	1.50	1.19	1.05	0.87	0.80	0.70

NOTES: The reported numbers are the average mean square errors (MSE) between model implied and quoted (bid-ask average) option prices relative to the average option price in a given maturity class, in percent. Gauss-Hermite expansion ($\alpha = 0.5$, $\beta = 1$) truncated after $M = 10 - 40$ terms, FIT method. The averages are computed for option data on June 19th, 2015 and across option maturities $\tau < 60$, $\tau \in (60, 180]$ and $\tau > 180$ days to expiration.

parity:

$$\widetilde{C}(K, \tau, S_t) = P(K, \tau, S_t) + e^{-q\tau} \cdot S_t - e^{-r\tau} \cdot K.$$

The put option prices are then removed from the sample. In such a way, for each observation date and option maturity we obtain the whole cross section of call option prices inferred from liquid ATM and OTM calls and puts.[23]

4.3 Case study: Triple witching on June 19th, 2015

We illustrate the calibration procedure by the example of June 19th, 2015. This is a triple witching date, i.e., one of the four dates in a year, when the contracts for stock index futures, stock index options and stock options expire simultaneously. These dates are associated with escalated trading activity and market liquidity, as traders close, roll out or offset their expiring positions. We extend the analysis for the whole sample of option data in next Section 4.4.

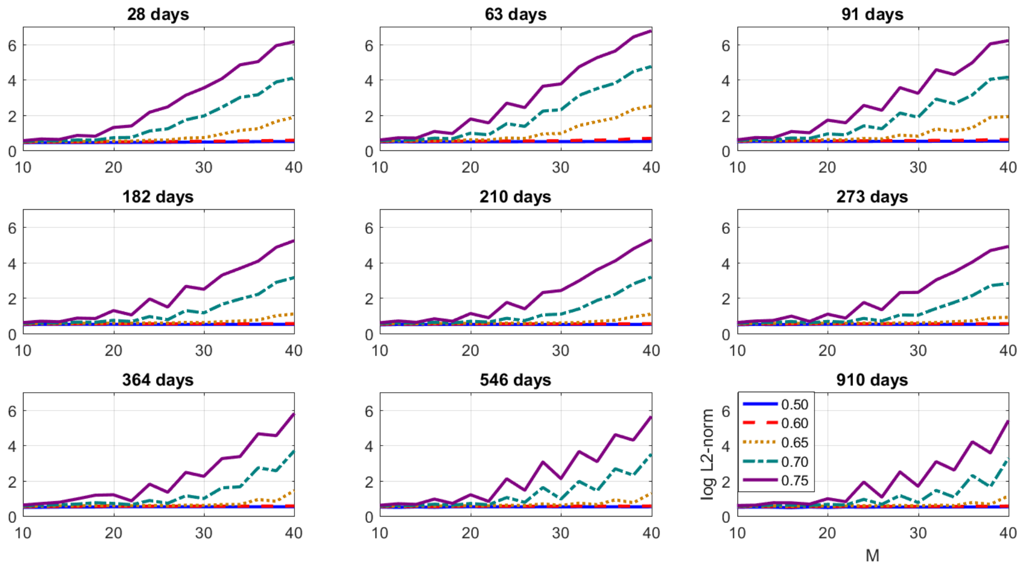
Table 4.2 reports the average mean squared errors (MSE) between model implied and quoted option prices for the Gauss-Hermite expansion truncated after $M = 10-40$ terms. The averages are computed across option maturities $\tau < 60$, $\tau \in (60, 180]$ and $\tau > 180$ days to expiration. We observe convergence of MSEs for short ($\tau \leq 60$) and medium ($\tau \in (60, 180]$) option maturities with the average MSE for $M = 40$ of less than 1% of the average call option price in the corresponding maturity class. Convergence of MSEs for long option maturities ($\tau > 180$) has not fully established by $M = 40$, but the model already explains market option prices well with the average MSE of less than 0.6% of the average call option price in the class.

Figure 4.1 illustrates the $\mathcal{L}^2_\alpha$-norms of approximating polynomials for α in $[0.5, 1]$ in logarithmic scale.[24] We observe that for $\alpha = 0.5$ (the Gauss-Hermite expansion)

[23] Even though we discard the data on illiquid options, the information embedded in these options partially transfers into the final dataset through continuously compounded dividend yields used in the call-put parity. Continuously compounded dividend yields q reported by Option Metrics are calculated using three months of call and put option data across all strikes and expirations.

[24] Recall that approximating polynomials for $\alpha > 0.5$ are obtained by expansing the Gauss-

Figure 4.1: $\mathcal{L}^2_\alpha$-norms of approximating polynomials on June 19th, 2015.



NOTES: The plotted values are the $\mathcal{L}^2_\alpha$-norms (4.4) of approximating polynomials implied by S&P 500 options on June 19th, 2015 in logarithmic scale. FIT method, $\beta = 1$, $\alpha = 0.5, 0.6, 0.65, 0.7$ and 0.75. Option maturities of 28–910 days.

the norms of approximating polynomials converge. Calibrating the model to higher orders M of approximating polynomials does not increase their norms significantly, which signals convergence of approximating norms to a finite number. On the other hand, for higher values of α, such as $\alpha = 0.7$ or $\alpha = 0.75$, we see no convergence of approximating norms within the given range of M. The norms of approximating polynomials increase exponentially with M.[25]

Figure 4.2 exhibits exponential growths of $\mathcal{L}^2_\alpha$-norms of approximating polynomials for $\alpha \in [0.5, 1]$. The growth rates $g^{(\alpha)}_{t,\tau}$ (for given observation date t and option maturity τ) are estimated from the linear model:

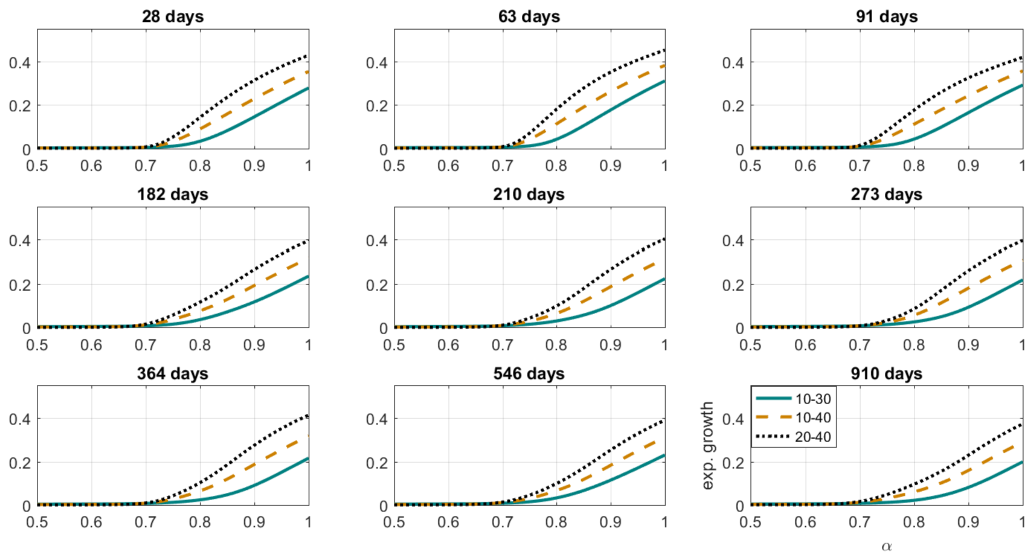
$$\log(S^{(\alpha)}_{t,\tau}(M)) = a + g^{(\alpha)}_{t,\tau} \cdot M + \varepsilon_{t,\tau,M} \tag{4.5}$$

with $S^{(\alpha)}_{t,\tau}(M)$ the $\mathcal{L}^2_\alpha$-norms of approximating polynomials defined in (4.4), $\varepsilon_{t,\tau,M}$ independent standard normal random variables and polynomial orders M in the range of $10 - 30$, $10 - 40$ and $20 - 40$. Figure 4.2 suggests that for each option maturity there is a threshold value $\alpha^* = \alpha^*(\tau)$, such that for every $\alpha > \alpha^*$ the exponential growth rates of approximating norms significantly differ from zero, i.e.,

Hermite approximating polynomials in the corresponding basis of generalized Hermite polynomials.

[25] Of course, our inferences on convergence are subject to the highest order ($M = 40$) of approximating polynomials used in calibration. It might be that after increasing M to, e.g., 100, we would observe convergence of approximating norms for $\alpha = 0.75$ or, alternatively, divergence of approximating norms for $\alpha = 0.55$.

Figure 4.2: Exponential growths of $\mathcal{L}^2_\alpha$-norms of approximating polynomials on June 19th, 2015.



NOTES: Exponential growths of $\mathcal{L}^2_\alpha$-norms (4.4) of approximating polynomials implied by S&P 500 options on June 19th, 2015. FIT method, $\beta = 1$. The growths are estimated from the linear model (4.5) with $M = 10 - 30$, $10 - 40$ and $20 - 40$ and for option maturities of 28–910 days.

the $\mathcal{L}^2_\alpha$-convergence of corresponding approximating polynomials fails. The pattern is robust to the range of Ms employed in estimation of exponential growths.

Since the $\mathcal{L}^2$-convergence of Gauss-Hermite approximating polynomials is well established, see, e.g., Myller-Lebedeff (1907), we employ the exponential growths of Gauss-Hermite approximating norms as the benchmark for empirical convergence of approximating norms for other $\alpha > 0.5$. Table 4.3 outputs these average growths as well as their average standard errors estimated from $M = 10-30$, $10-40$ and $20-40$. As expected, we find that for both short, medium and long option maturities the exponential growths decrease with the range of Ms, which implies establishment of convergence of the Gauss-Hermite approximating norms.

Finally, Table 4.4 reports the thresholds α^* and the corresponding tail parameters $\hat{\xi} = 1 - \frac{1}{2\alpha^*}$ (formula (4.2)), based on the empirical $\mathcal{L}^2_\alpha$-convergence of approximating polynomials benchmarked by 1, 2, and 3 average standard deviations of Gauss-Hermite exponential growths (over all option maturities) in excess to the Gauss-Hermite exponential growth for a given option maturity. We find that for all option maturities on June 19th, 2015, the tail parameter $\hat{\xi}$ is in the range of $[0.15, 0.35]$ (vs $\xi = 1$ for normally distributed tails of the RND). Accordingly, the tails of the model implied RND are heavier than normally distributed, but are not fat, as the parameter $\hat{\xi}$ never reaches zero.

We note that the tail parameter $\hat{\xi}$ increases (implying lighter tails) with the

Table 4.3: Exponential growths of $\mathcal{L}^2$-norms of Gauss-Hermite approximating polynomials on June 19th, 2015.

Days-to-exp.	≤ 60			$(60, 180]$			> 180			all		
M	10-30	10-40	20-40	10-30	10-40	20-40	10-30	10-40	20-40	10-30	10-40	20-40
av. gr., 10^{-3}	2.15	1.74	1.28	3.49	2.21	0.90	4.27	2.62	0.94	3.86	2.43	0.97
av. std., 10^{-3}	0.56	0.33	0.14	0.81	0.53	0.09	0.92	0.59	0.17	0.85	0.55	0.15

NOTES: The reported numbers are average exponential growths and their average Newey-West standard errors estimated from the linear model (4.5) with $\alpha = 0.5$ and $M = 10 - 30$, $10 - 40$ and $20 - 40$. The averages are computed for option data on June 19th, 2015 and across option maturities $\tau < 60$, $\tau \in (60, 180]$ and $\tau > 180$ days to expiration.

convergence benchmark (here 1, 2, or 5 std) and decreases (implying heavier tails) when polynomials of higher orders are used in the estimation procedure. Indeed, the closer the convergence benchmark (here, 1, 2, or 5 std) to the exponential growth of Gauss-Hermite approximating norms, the smaller the identified threshold α^*, i.e., the heavier the tails of the model implied RND. Monotonicity of the tail parameter on the highest polynomial order M results from $\mathcal{L}^2$-convergence establishing for higher orders of M only. It is unclear, which of the two effects is stronger, implying that the estimated tail parameters might differ quantitatively, but also qualitatively (if $\hat{\xi} = 0$) if the model were possible to calibrate for $M > 40$.

4.4 Time series of tail parameters

We now apply the procedure of Section 4.3 to recover the time-series of tail parameters (4.2) between January 1st, 2014 and December 31st, 2015.

Tables 4.5 and 4.6, respectively, report the average mean square errors of the Gauss-Hermite expansion and the average exponential growths with standard deviations of the $\mathcal{L}^2$-norms of Gauss-Hermite approximating polynomials. The results are qualitatively the same as for the case study on June 19th, 2015.

Figure 4.3 plots the time-series of tail parameters (4.2), based on the empirical $\mathcal{L}^2_\alpha$-convergence of approximating polynomials benchmarked by 2 and 5 average standard deviations of Gauss-Hermite exponential growths (for all observation dates and option maturities) in excess to Gauss-Hermite exponential growths for each given observation date and option maturity. The exponential growths are estimated with $M = 10 - 40$ and $M = 20 - 40$. The tail parameter is averaged within a class of short ($\tau \leq 60$), medium ($\tau \in (60, 180]$) and long ($\tau > 180$) option maturities.

We observe that for all observation dates and all option maturities the tail parameter is within the range of $[0.1, 0.3]$, which signals heavier than normally distributed tails of the model implied RND, but no fat tails. The convergence benchmark (here, 2 or 5 std) and the range of M employed in estimation of exponential growths have

Table 4.4: Threshold α^*s and tail parameters $\hat{\xi}$ on June 19th, 2015.

conv. ben. (in std)	1						2						5					
M	10–30		10–40		20–40		10–30		10–40		20–40		10–30		10–40		20–40	
$\tau\backslash$ stat.	α^*	$\hat{\xi}$	α^*	$\hat{\xi}$	α^*	$\hat{\xi}$	α^*	$\hat{\xi}$	α^*	$\hat{\xi}$	α^*	$\hat{\xi}$	α^*	$\hat{\xi}$	α^*	$\hat{\xi}$	α^*	$\hat{\xi}$
28 d.	0.64	0.219	0.66	0.242	0.66	0.242	0.69	0.275	0.68	0.265	0.67	0.254	0.73	0.315	0.70	0.286	0.68	0.265
63 d.	0.61	0.180	0.61	0.180	0.61	0.180	0.68	0.265	0.66	0.242	0.64	0.219	0.73	0.315	0.69	0.275	0.67	0.254
91 d.	0.61	0.180	0.61	0.180	0.59	0.153	0.67	0.254	0.66	0.242	0.64	0.219	0.72	0.306	0.68	0.265	0.66	0.242
182 d.	0.60	0.167	0.59	0.153	0.56	0.107	0.66	0.242	0.63	0.206	0.60	0.167	0.71	0.296	0.67	0.254	0.63	0.206
210 d.	0.60	0.167	0.59	0.153	0.56	0.107	0.65	0.231	0.64	0.219	0.60	0.167	0.71	0.296	0.67	0.254	0.64	0.219
273 d.	0.60	0.167	0.59	0.153	0.55	0.091	0.65	0.231	0.64	0.219	0.59	0.153	0.71	0.296	0.68	0.265	0.64	0.219
364 d.	0.61	0.180	0.60	0.167	0.55	0.091	0.66	0.242	0.64	0.219	0.60	0.167	0.72	0.306	0.68	0.265	0.64	0.219
546 d.	0.60	0.167	0.59	0.153	0.55	0.091	0.66	0.242	0.63	0.206	0.58	0.138	0.71	0.296	0.66	0.242	0.62	0.194
910 d.	0.60	0.167	0.59	0.153	0.57	0.123	0.66	0.242	0.63	0.206	0.60	0.167	0.72	0.306	0.66	0.242	0.62	0.194

NOTES: The reported numbers are the threshold values (4.3) and the corresponding tail parameters (4.2) implied by S&P 500 options on June 19th, 2015 with maturities of 28–910 days to expiration. The tail parameters are estimated based on the empirical $\mathcal{L}^2_\alpha$-convergence of approximating polynomials benchmarked by 1, 2 and 5 average standard deviations of Gauss-Hermite exponential growths (across all option maturities) in excess to the Gauss-Hermite exponential growth for a given option maturity. Exponential growths are computed with $M = 10 - 30$, $M = 10 - 40$ and $M = 20 - 40$.

Table 4.5: MSE of the Gauss-Hermite expansion in 2014-2015.

Days-to-exp.\ M	10	12	14	16	18	20	22	24	26	28	30	32	34	36	38	40
≤ 60	6.50	4.13	3.52	2.91	2.05	1.65	1.52	1.33	1.24	1.12	1.04	0.97	0.93	0.88	0.84	0.82
$(60, 180]$	10.35	6.28	3.97	2.89	2.18	1.41	1.14	0.78	0.68	0.52	0.48	0.40	0.37	0.33	0.32	0.29
> 180	35.65	17.35	12.99	6.19	6.28	3.61	3.60	2.49	2.04	1.50	1.26	0.97	0.85	0.71	0.64	0.56
all	23.73	12.05	8.92	4.74	4.45	2.69	2.58	1.84	1.55	1.18	1.02	0.82	0.74	0.65	0.59	0.54

NOTES: The reported numbers are the average mean square errors (MSE) between model implied and quoted (bid-ask average) option prices relative to the average option price in a given maturity class, in percent. Gauss-Hermite expansion ($\alpha = 0.5$, $\beta = 1$) truncated after $M = 10 - 40$ terms, FIT method. The averages are computed for the sample period from January 1st, 2014 to December 31st, 2015 and across option maturities $\tau < 60$, $\tau \in (60, 180]$ and $\tau > 180$ days to expiration.

Table 4.6: Exponential growths of $\mathcal{L}^2$-norms of Gauss-Hermite approximating polynomials in 2014-2015.

Days-to-exp.	≤ 60			$(60, 180]$			> 180			all		
M	10-30	10-40	20-40	10-30	10-40	20-40	10-30	10-40	20-40	10-30	10-40	20-40
av. gr., 10^{-3}	3.80	2.52	1.23	4.22	2.61	0.96	4.43	2.74	1.02	4.26	2.66	1.04
av. std, 10^{-3}	0.73	0.41	0.13	0.89	0.48	0.14	0.95	0.62	0.21	0.89	0.55	0.18

NOTES: The reported numbers are the average exponential growths and their average Newey-West standard errors estimated from the linear model (4.5) with $\alpha = 0.5$ and $M = 10 - 30$, $10 - 40$ and $20 - 40$. The averages are computed for the sample period from January 1st, 2014 to December 31st, 2015 and across option maturities $\tau < 60$, $\tau \in (60, 180]$ and $\tau > 180$ days to expiration.

the same effect as in the case study on June 19th, 2015.

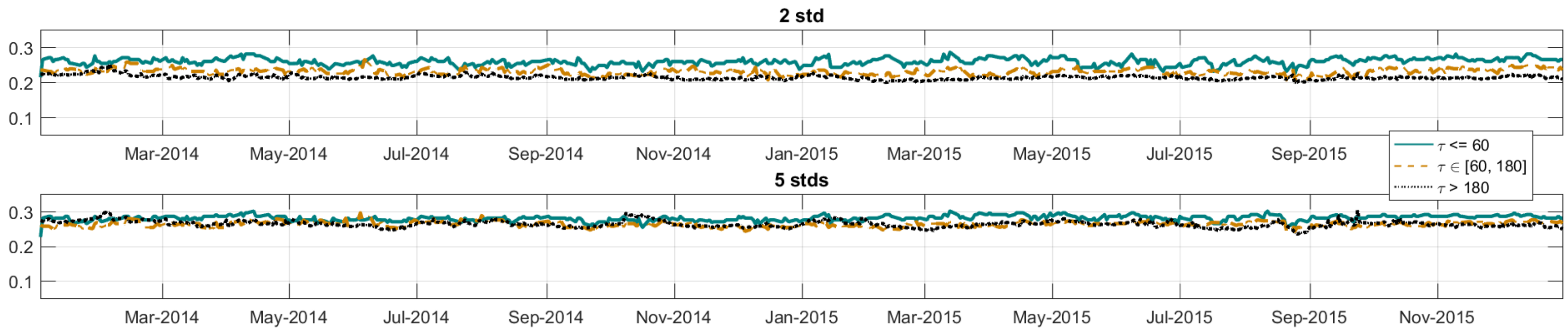
Figure 4.4 also plots the tail parameters estimated using the expansion coefficients calibrated with the IV method (Section 3.1.1) instead of the FIT method. We observe that the IV method consistently underestimates the tail parameter relative to the FIT method, i.e., implies heavier tails of approximating RND.[26] Nevertheless, even when calibrated with the IV method, the model never detects presence of fat tails in the risk-neutral density implied by S&P 500 options in 2014–2015.

5 Conclusion and outlook

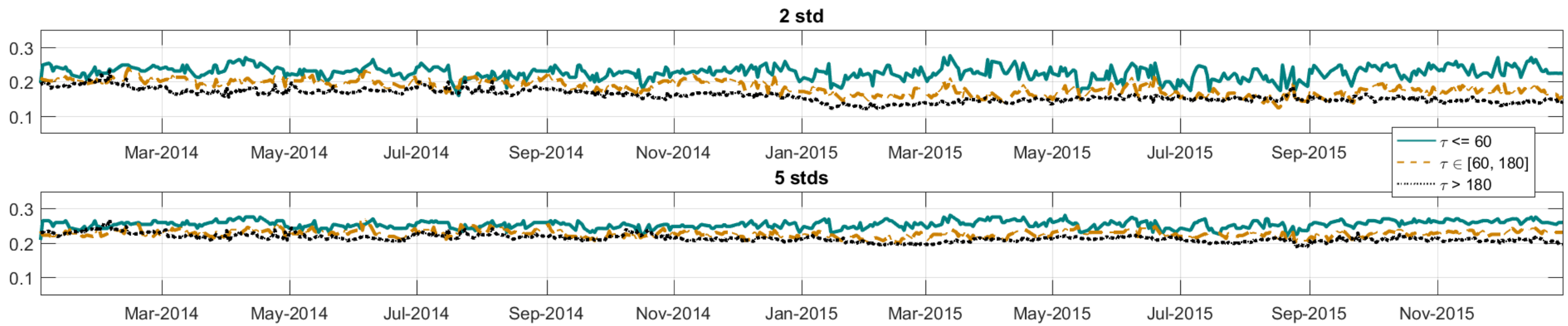
Approximating the risk-neutral density using the generalized Hermite expansion yields a closed form option pricing formula as given in Theorem 3.1. The formula embeds the classical Black and Scholes (1973) formula, the option pricing formula based on the Gram-Charlier Type A expansion as in Corrado (2007) and the option

[26] As discussed in Section 3.1.1, the inconsistency results from truncation and discretization errors in formula (3.2).

Figure 4.3: Tail parameter (4.2) in 2014-2015, FIT method.



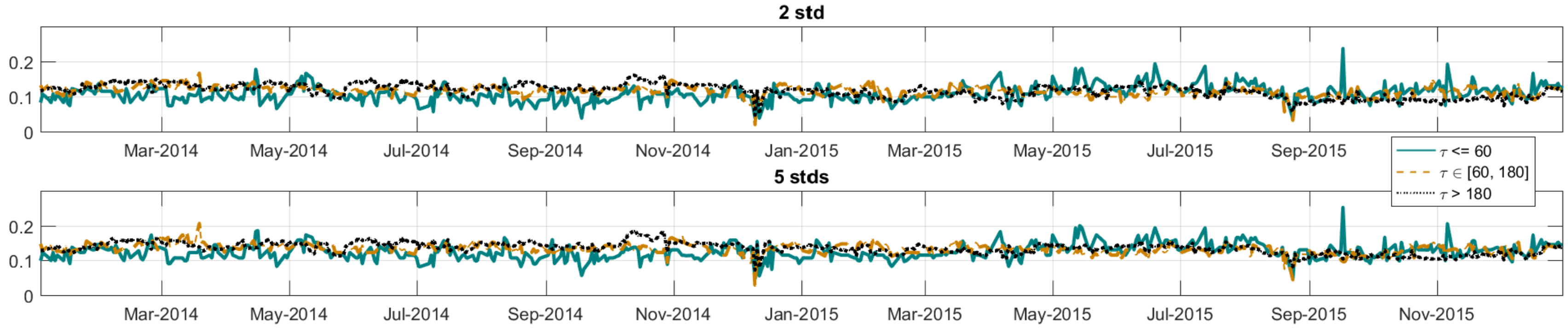
(a) $M = 10 - 40$



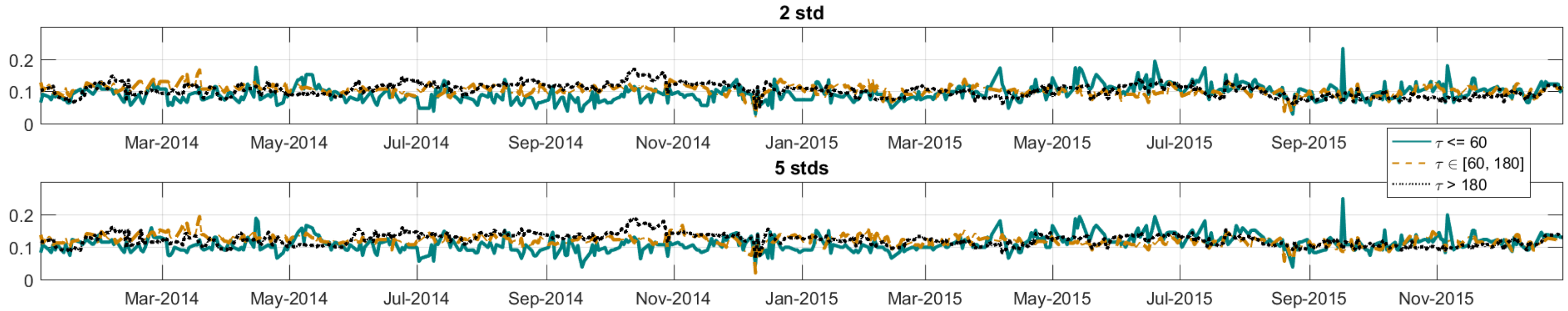
(b) $M = 20 - 40$

NOTES: Time series of tail parameters (4.2) for the sample period from January 1st, 2014 to December 31st, 2015, FIT method. The plotted values are the average tail parameters across option maturities $\tau < 60$, $\tau \in (60, 180]$ and $\tau > 180$ days to expiration. Tail parameters are estimated based on the empirical $\mathcal{L}^2_\alpha$-convergence of approximating polynomials benchmarked by 2 and 5 average standard deviations of Gauss-Hermite exponential growths (across all observation dates and option maturities) in excess to the Gauss-Hermite exponential growth for a given observation date and option maturity. $M = 10 - 40$ (a) and $M = 20 - 40$ (b) refer to the range of approximating polynomials employed in computing exponential growths of $\mathcal{L}^2_\alpha$-norms.

Figure 4.4: Tail parameter (4.2) in 2014-2015, IV method.



(a) $M = 10 - 40$



(b) $M = 20 - 40$

NOTES: Time series of tail parameters (4.2) for the sample period from January 1st, 2014 to December 31st, 2015, IV method. The plotted values are the average tail parameters across option maturities $\tau < 60$, $\tau \in (60, 180]$ and $\tau > 180$ days to expiration. Tail parameters are estimated based on the empirical $\mathcal{L}^2_\alpha$-convergence of approximating polynomials benchmarked by 2 and 5 average standard deviations of Gauss-Hermite exponential growths (across all observation dates and option maturities) in excess to the Gauss-Hermite exponential growth for a given observation date and option maturity. $M = 10 - 40$ (a) and $M = 20 - 40$ (b) refer to the range of approximating polynomials employed in computing exponential growths of $\mathcal{L}^2_\alpha$-norms.

pricing formula based on the Gauss-Hermite expansion obtained recently in Necula et al. (2016).

We derived four alternative methods for obtaining the expansion coefficients. Specifically, one can obtain the coefficients of the generalized Hermite expansion from the probability distribution function using formula (2.6), from the characteristic function using the results in Proposition 2.2, or calibrate them to market option prices using either results in Proposition 3.2 (IV method) or by minimizing the total square error between model implied and quoted option prices (FIT method). Unlike the particular case of Edgeworth expansions studied previously in the literature, a key advantage of the generalized Hermite expansion is its applicability to heavy-tailed return distributions, an aspect of key importance in option markets.

We also employed the generalized Hermite expansion to analyse convergence of the tails of the RND implied by European options on the S&P 500 price index. Subject to the estimation procedure, we found that the RND with tails heavier than of normal distribution but not fat, can explain the market option data well.

Our quantitative results on the tails of options implied RND depend on the estimation methodology, hence should not be misinterpreted as the true information on the RND tails embedded in option prices. Firstly, we could not increase the highest polynomial order M beyond 40 because of computational issues arising. Secondly, the estimated tail parameter relies on the convergence benchmark, which ideally should be set as close as possible to zero. Thirdly, our option data is limited to minimum and maximum traded strike price, whereas extending strikes interval may potentially impact tails behaviour of the model implied RND.

Given the above, it would be interesting to further explore the effects of the highest polynomial order, the convergence benchmark and the strikes interval on the tails of the model implied RND. Moreover, one could employ the generalized Hermite expansion to alternative derivatives and investigate whether the qualitative results on the tails of the RND would hold in the general case too.

References

Abken, P. A., D. B. Madan, and S. Ramamurtie (1996a). Estimation of risk-neutral and statistical densities by Hermite polynomial approximation: With an application to Eurodollar futures options. Technical report, Federal Reserve Bank of Atlanta.

Abken, P. A., D. B. Madan, and S. Ramamurtie (1996b). Pricing S&P 500 index options using a Hilbert space basis. Technical report, Federal Reserve Bank of Atlanta.

Aït-Sahalia, Y. and J. Duarte (2003). Nonparametric option pricing under shape restrictions. *Journal of Econometrics 116*(1), 9–47.

Aït-Sahalia, Y. and A. W. Lo (1998). Nonparametric estimation of state-price densities implicit in financial asset prices. *The Journal of Finance 53*(2), 499–547.

Amin, K. I. and R. A. Jarrow (1992). Pricing options on risky assets in a stochastic interest rate economy. *Mathematical Finance 2*(4), 217–237.

Andersen, T. G., N. Fusari, and V. Todorov (2015). The risk premia embedded in index options. *Journal of Financial Economics 117*(3), 558–584.

Andrews, G. E., R. Askey, and R. Roy (1999). Special functions. 1999. *Encyclopedia Math. Appl.*

Bakshi, G., N. Kapadia, and D. Madan (2003). Stock return characteristics, skew laws, and the differential pricing of individual equity options. *Review of Financial Studies 16*(1), 101–143.

Bates, D. S. (1991). The crash of'87: Was it expected? The evidence from options markets. *The Journal of Finance 46*(3), 1009–1044.

Bates, D. S. (1996). Jumps and stochastic volatility: Exchange rate processes implicit in deutsche mark options. *Review of Financial Studies 9*(1), 69–107.

Bates, D. S. (2003). Empirical option pricing: A retrospection. *Journal of Econometrics 116*(1), 387–404.

Bibby, B. M. and M. Sørensen (1996). A hyperbolic diffusion model for stock prices. *Finance and Stochastics 1*(1), 25–41.

Birru, J. and S. Figlewski (2012). Anatomy of a meltdown: The risk neutral density for the S&P 500 in the fall of 2008. *Journal of Financial Markets 15*(2), 151–180.

Black, F. and M. Scholes (1973). The pricing of options and corporate liabilities. *Journal of Political Economy 81*(3), 637–654.

Bollerslev, T. and V. Todorov (2011). Tails, fears, and risk premia. *The Journal of Finance 66*(6), 2165–2211.

Breeden, D. T. and R. H. Litzenberger (1978). Prices of state-contingent claims implicit in option prices. *Journal of Business*, 621–651.

Brenner, M. and Y. H. Eom (1997). No-arbitrage option pricing: New evidence on the validity of the martingale property. Working paper.

Carr, P. and L. Wu (2009). Variance risk premiums. *Review of Financial Studies 22*(3), 1311–1341.

Chen, R.-R., P.-l. Hsieh, J. Huang, and K. Bank (2017). Crash risk and risk neutral densities. Working paper.

Corrado, C. (2007). The hidden martingale restriction in Gram-Charlier option prices. *Journal of Futures Markets 27*(6), 517–534.

Corrado, C. J. and T. Su (1996). Skewness and kurtosis in S&P 500 index returns implied by option prices. *Journal of Financial Research 19*(2), 175–192.

Corrado, C. J. and T. Su (1997). Implied volatility skews and stock index skewness and kurtosis implied by S&P 500 index option prices. *The Journal of Derivatives 4*(4), 8–19.

Corrado, C. J., T. Su, et al. (1996). S&P 500 index option tests of Jarrow and Rudd's approximate option valuation formula. *Journal of Futures Markets 16*(6), 611–629.

Filipović, D., E. Mayerhofer, and P. Schneider (2013). Density approximations for multivariate affine jump-diffusion processes. *Journal of Econometrics 176*(2), 93–111.

Hamidieh, K. (2017). Estimating the tail shape parameter from option prices. *Journal of Risk 19*(6), 85–110.

Hens, T. and K. R. Schenk-Hoppé (2009). *Handbook of financial markets: Dynamics and evolution*. Elsevier.

Heston, S. L. (1993). A closed-form solution for options with stochastic volatility with applications to bond and currency options. *Review of Financial Studies 6*(2), 327–343.

Hommes, C. (2013). *Behavioral Rationality and Heterogeneous Expectations in Complex Economic Systems*. Cambridge University Press.

Hull, J. and A. White (1987). The pricing of options on assets with stochastic volatilities. *The Journal of Finance 42*(2), 281–300.

Jackwerth, J. C. (2000). Recovering risk aversion from option prices and realized returns. *Review of Financial Studies 13*(2), 433–451.

Jarrow, R. and A. Rudd (1982). Approximate option valuation for arbitrary stochastic processes. *Journal of Financial Economics 10*(3), 347–369.

Jiang, G. J. and Y. S. Tian (2005). The model-free implied volatility and its information content. *Review of Financial Studies 18*(4), 1305–1342.

Jurczenko, E., B. Maillet, and B. Negréa (2004). A note on skewness and kurtosis adjusted option pricing models under the martingale restriction. *Quantitative Finance 4*(5), 479–488.

Kaeck, A. and C. Alexander (2012). Volatility dynamics for the S&P 500: Further evidence from non-affine, multi-factor jump diffusions. *Journal of Banking and Finance 36*(11), 3110–3121.

Kou, S. G. (2002). A jump-diffusion model for option pricing. *Management Science 48*(8), 1086–1101.

Madan, D. B., P. P. Carr, and E. C. Chang (1998). The variance gamma process and option pricing. *European Finance Review 2*(1), 79–105.

Madan, D. B. and F. Milne (1994). Contingent claims valued and hedged by pricing and investing in a basis. *Mathematical Finance 4*(3), 223–245.

Malevergne, Y., V. Pisarenko, and D. Sornette (2006). On the power of generalized extreme value (GEV) and generalized Pareto distribution (GPD) estimators for empirical distributions of stock returns. *Applied Financial Economics 16*(3), 271–289.

Merton, R. C. (1973). Theory of rational option pricing. *The Bell Journal of Economics and Management Science*, 141–183.

Merton, R. C. (1976). Option pricing when underlying stock returns are discontinuous. *Journal of Financial Economics 3*(1-2), 125–144.

Myller-Lebedeff, W. (1907). Die theorie der integralgleichungen in anwendung auf einige reihenentwicklungen. *Mathematische Annalen 64*(3), 388–416.

Necula, C., G. Drimus, and W. Farkas (2016). A general closed form option pricing formula. Working paper.

Orosi, G. (2015). Information content of right option tails: Evidence from S&P 500 index options. *Journal of Asset Management*, 1–11.

Rompolis, L. S. and E. Tzavalis (2007). Retrieving risk neutral densities based on risk neutral moments through a Gram–Charlier series expansion. *Mathematical and Computer Modelling 46*(1), 225–234.

Scott, L. O. (1997). Pricing stock options in a jump-diffusion model with stochastic volatility and interest rates: Applications of Fourier inversion methods. *Mathematical Finance 7*(4), 413–426.

Stein, E. M. and J. C. Stein (1991). Stock price distributions with stochastic volatility: An analytic approach. *Review of Financial Studies 4*(4), 727–752.

Vilkov, G. and Y. Xiao (2013). Option-implied information and predictability of extreme returns. Working paper.

Appendix

A Proofs of the main results

Lemma A.1 *If $f(x)$ is uniformly continuous on $x \in [a, \infty)$ and $\int_a^\infty f(x)dx$ converges, then*

$$\lim_{x\to\infty} f(x) = 0.$$

Proof of Lemma A.1: Consider $\varepsilon > 0$. Since f is uniformly continuous, there exists $\delta = \delta(\varepsilon) \in (0, 1)$, such that

$$|f(x) - f(y)| < \varepsilon/2 \text{ for any } x, y \text{ with } |x - y| < \delta.$$

On the other hand, since

$$\int_a^\infty f(x)dx = \int_a^{([a/\delta]+1)\delta} f(x)dx + \sum_{k=[a/\delta]+1}^{\infty} \int_{k\delta}^{(k+1)\delta} f(x)dx$$

converges, we have $\lim_{k\to\infty} \int_{k\delta}^{(k+1)\delta} f(x)dx = 0$. The latter implies that there exists $N > 0$, such that $\left|\int_{k\delta}^{(k+1)\delta} f(x)dx\right| < \frac{\varepsilon\delta}{2}$ for all $k \geq N$.

Consider now $x_0 \geq (N + 1)\delta$. There exists $k \geq N$, such that $x_0 \in [k\delta, (k + 1)\delta]$. By uniform continuity it follows that $|f(x) - f(x_0)| < \varepsilon/2$ for all $x \in [k\delta, (k + 1)\delta]$. Therefore,

$$\left|\int_{k\delta}^{(k+1)\delta} f(x)dx\right| \geq \delta(|f(x_0)| - \varepsilon/2).$$

On the other hand, since $\left|\int_{k\delta}^{(k+1)\delta} f(x)dx\right| < \frac{\varepsilon\delta}{2}$, the latter implies that $|f(x_0)| < \varepsilon$. Consequently, $\lim_{x\to\infty} f(x) = 0$, completing the proof. □

A.1 Characteristic function representation

Proof of Proposition 3.2: By definition of characteristic function we have:

$$\varphi_\mathcal{T}(u) = \int_\mathbb{R} e^{iux} p_\mathcal{T}(x)dx = \int_\mathbb{R} e^{iu(\mu_\mathcal{T}+\sigma_\mathcal{T} y)} p_\mathcal{T}(\mu_\mathcal{T} + \sigma_\mathcal{T} y)\sigma_\mathcal{T} dy = e^{iu\mu_\mathcal{T}} \int_\mathbb{R} e^{iu\sigma_\mathcal{T} y} \widetilde{p}_\mathcal{T}(y)dy,$$

where $\widetilde{p}_\mathcal{T}$ is the standardized risk-neutral density function, and the second equality holds by the change of variable $x = \mu_\mathcal{T} + \sigma_\mathcal{T} y$.

Denote by $R_m^{(\alpha,\beta)}(x)$ the partial sums of the generalized Hermite expansion with

$$R_m^{(\alpha,\beta)}(x) := \sum_{n=0}^{m} a_{\mathcal{T},n}^{(\alpha,\beta)} \cdot H_n^{(\alpha)}(x). \tag{A.1}$$

Recall that $\widetilde{p}_\mathcal{T}(x) = \lim_{m\to\infty} z^{(\beta)}(x) \cdot R_m^{(\alpha,\beta)}(x)$ with convergence holding in $\mathcal{L}^2_\alpha$. Let us now show that $\lim_{m\to\infty} \Delta_m = 0$, where

$$\Delta_m := \lim_{m\to\infty} \int_{\mathbb{R}} e^{iu\sigma_\tau y} \big[\widetilde{p}_\tau(y) - R_m^{(\alpha,\beta)}(y) \cdot z^{(\beta)}(y)\big],$$

which will allow to compute the integral in $\varphi_\tau(u)$ term by term. We have

$$|\Delta_m| \le \int_{\mathbb{R}} \left| \frac{\widetilde{p}_\tau(y)}{z^{(\beta)}(y)} - R_m^{(\alpha,\beta)}(y) \right| z^{(\beta)}(y)dy \le \left|\left| \frac{\widetilde{p}_\tau(y)}{z^{(\beta)}(y)} - R_m^{(\alpha,\beta)}(y) \right|\right|_{\mathcal{L}^2_\alpha} \cdot \left|\left| \frac{z^{(\beta)}(y)}{w^{(\alpha)}(y)} \right|\right|_{\mathcal{L}^2_\alpha},$$

where the second inequality is the Hoelder's inequality with $p = q = 2$. The left norm converges to 0 as $m \to \infty$. The right norm is finite with

$$\left|\left| \frac{z^{(\beta)}(y)}{w^{(\alpha)}(y)} \right|\right|^2_{\mathcal{L}^2_\alpha} = \int_{\mathbb{R}} \left[\frac{z^{(\beta)}(x)}{w^{(\alpha)}(x)} \right]^2 w^{(\alpha)}(x)dx = \frac{1}{2\pi\beta} \int_{\mathbb{R}} e^{-\frac{x^2}{2}\left(\frac{2}{\beta} - \frac{1}{\alpha}\right)} dx < \infty$$

for any $\alpha > \frac{\beta}{2}$. This implies that $\lim_{m\to\infty} \Delta_m = 0$ does hold.

Consequently, we can compute the characteristic function as

$$\varphi_\tau(u) = e^{iu\mu_\tau} \cdot \sum_{n=0}^{\infty} a_{\tau,n}^{(\alpha,\beta)} \cdot \int_{\mathbb{R}} e^{iu\sigma_\tau x} \cdot z^{(\beta)}(x) \cdot H_n^{(\alpha)}(x)dx := e^{iu\mu_\tau} \cdot \sum_{n=0}^{\infty} a_{\tau,n}^{(\alpha,\beta)} \cdot L_n(u), \tag{A.2}$$

where $L_n(u)$ denote the integral terms. It is straightforward to see that differentiation with respect to u under the integral sign in $L_n(u)$ is allowed and leads to

$$L_n'(u) = i\sigma_\tau \int_{\mathbb{R}} x e^{iu\sigma_\tau x} \cdot z^{(\beta)}(x) \cdot H_n^{(\alpha)}(x)dx.$$

Using the recursive property (2.1), the differentiation property (2.3) as well as integration by parts, we derive a recursive relationship for the sequence $L_n(u)$:

$$\begin{aligned} L_n(u) &= \int_{\mathbb{R}} e^{iu\sigma_\tau x} \cdot z^{(\beta)}(x) \cdot \left(\frac{x}{\alpha} H_{n-1}^{(\alpha)}(x) - H_{n-1}^{(\alpha)\prime}(x) \right) dx \\ &= \frac{1}{i\sigma_\tau \alpha} \cdot L_{n-1}'(u) + \int_{\mathbb{R}} e^{iu\sigma_\tau x} \left(iu\sigma_\tau \cdot z^{(\beta)}(x) - \frac{x}{\beta} z^{(\beta)}(x) \right) H_{n-1}^{(\alpha)}(x)dx \\ &= iu\sigma_\tau \cdot L_{n-1}(u) + \frac{1}{i\sigma_\tau} \cdot \frac{\beta - \alpha}{\alpha\beta} \cdot L_{n-1}'(u). \end{aligned} \tag{A.3}$$

Finally, we let $L_n(u) = i^n \cdot e^{-\frac{u^2\sigma_\tau^2\beta}{2}} \cdot G_n^{(\alpha,\beta)}(\sigma_\tau u)$. Upon differentiation this yields to

$$L_n'(u) = i^n \cdot e^{-\frac{u^2\sigma_\tau^2\beta}{2}} \cdot \big(-u \cdot \sigma_\tau^2 \cdot \beta \cdot G_n^{(\alpha,\beta)}(u\sigma_\tau) + \sigma_\tau \cdot G_n^{(\alpha,\beta)\prime}(u\sigma_\tau) \big).$$

Inserting the latter into the recursive relationship (A.3) and rearranging the terms gives:

$$G_n^{(\alpha,\beta)}(u\sigma_\tau) = u \cdot \sigma_\tau \cdot \frac{\beta}{\alpha} \cdot G_{n-1}^{(\alpha,\beta)}(u\sigma_\tau) - \frac{\beta - \alpha}{\alpha\beta} \cdot G_{n-1}^{(\alpha,\beta)\prime}(u\sigma_\tau)$$

for any $u \in \mathbb{R}$, or equivalently,

$$G_n^{(\alpha,\beta)}(x) = x \cdot \frac{\beta}{\alpha} \cdot G_{n-1}^{(\alpha,\beta)}(x) - \frac{\beta-\alpha}{\alpha\beta} \cdot G_{n-1}^{(\alpha,\beta)\prime}(x).$$

It is straightforward to check that $G_0^{(\alpha,\beta)}(x) = 1$, $G_1^{(\alpha,\beta)}(x) = x \cdot \frac{\beta}{\alpha}$ and $G_n^{(\alpha,\beta)}(x)$ is a polynomial of degree n, which completes the proof. $\square$

A.2 Closed-form option pricing formula

Proof of Theorem 3.1: By the standard risk-neutral valuation approach we have:

$$\begin{aligned} C\big(K,\tau,S_t,r,q,\mu,\sigma,a_n^{(\alpha,\beta)}\big) &= e^{-r\tau}\int_{\log(K/S_t)}^{\infty}(S_t \cdot e^x - K)\cdot p_\tau(x)\cdot dx \\ &= e^{-r\tau}\int_{-d_2}^{\infty}(S_t \cdot e^{\mu\tau+\sigma\sqrt{\tau}y} - K)\cdot \widetilde{p}_\tau(y)\cdot dy \end{aligned}$$

with the second equality holding by the change of variable $x = \mu\tau + \sigma\sqrt{\tau}y$.

We compute the two terms in the expression above separately. Let us first check that $\lim_{m\to\infty}\Delta_m = 0$, where

$$\Delta_m := \lim_{t\to\infty}\int_{-d_2}^{\infty} e^{\sigma\sqrt{\tau}y}\cdot\big[\widetilde{p}_\tau(y) - R_m^{(\alpha,\beta)}(y)\cdot z^{(\beta)}(y)\big]dy$$

with $R_m^{(\alpha,\beta)}(x)$ defined in (A.1). We have:

$$\begin{aligned} |\Delta_m| &\le \int_{-d_2}^{\infty} e^{\sigma\sqrt{\tau}y}\left|\frac{\widetilde{p}_\tau(y)}{z^{(\beta)}(y)} - R_m^{(\alpha,\beta)}(y)\right| z^{(\beta)}(y)dy \\ &\le \left(\int_{-d_2}^{\infty}\left|\frac{\widetilde{p}_\tau(x)}{z^{(\beta)}(x)} - R_m^{(\alpha,\beta)}(x)\right|^2 w^{(\alpha)}(x)dx \cdot \int_{-d_2}^{\infty} e^{2\sigma\sqrt{\tau}x}\left[\frac{z^{(\beta)}(x)}{w^{(\alpha)}(x)}\right]^2 w^{(\alpha)}(x)dx\right)^{1/2} \\ &\le \left|\left|\frac{\widetilde{p}_\tau(x)}{z^{(\beta)}(x)} - R_m^{(\alpha,\beta)}(x)\right|\right|_{\mathcal{L}^2_\alpha} \cdot \varkappa_1^{(\alpha,\beta)} \end{aligned}$$

with $\varkappa_1^{(\alpha,\beta)} = \left(\frac{1}{2\pi\beta}\int_{-d_2}^{\infty} e^{2\sigma\sqrt{\tau}x} e^{-\frac{x^2}{2}\left(\frac{2}{\beta}-\frac{1}{\alpha}\right)}dx\right)^{1/2} < \infty$ for $\alpha > \frac{\beta}{2}$. Since $R_m^{(\alpha,\beta)}$ converge with $m \to \infty$ to $\frac{\widetilde{p}(y)}{z^{(\beta)}(y)}$ in $\mathcal{L}^2_\alpha$, this proves that condition $\lim_{m\to\infty}\Delta_m = 0$ does hold.

Consequently, we are allowed to interchange the integral and the infinite series to obtain:

$$e^{-r\tau}\int_{-d_2}^{\infty} S_t \cdot e^{\mu\tau+\sigma\sqrt{\tau}y}\cdot\widetilde{p}_\tau(y)\cdot dy = S_t \cdot e^{\left(\mu - r + \frac{\sigma^2\beta}{2}\right)\tau}\cdot\sum_{n=0}^{\infty} a_n^{(\alpha,\beta)}\cdot I_n^{(\alpha,\beta)},$$

where

$$\begin{aligned}
I_n^{(\alpha,\beta)} &:= e^{-\frac{\sigma^2\beta\tau}{2}} \int_{-d_2}^{\infty} e^{\sigma\sqrt{\tau}x} \cdot z^{(\beta)}(x) \cdot H_n^{(\alpha)}(x)dx \\
&= e^{-\frac{\sigma^2\beta\tau}{2}} \int_{-d_2-\sigma\beta\sqrt{\tau}}^{\infty} e^{\sigma\sqrt{\tau}(x+\sigma\beta\sqrt{\tau})} \cdot z^{(\beta)}(x+\sigma\beta\sqrt{\tau}) \cdot H_n^{(\alpha)}(x+\sigma\beta\sqrt{\tau})dx \\
&= \int_{-d_2-\sigma\beta\sqrt{\tau}}^{\infty} H_n^{(\alpha)}(x+\sigma\beta\sqrt{\tau}) z^{(\beta)}(x)dx.
\end{aligned}$$

To establish the recurrence relation for the sequence $I_n^{(\alpha)}$, we use the properties (2.1), (2.4) and integration by parts to obtain:

$$\begin{aligned}
I_{n+1}^{(\alpha,\beta)} &= \int_{-d_2-\sigma\beta\sqrt{\tau}}^{\infty} \left(\frac{x+\sigma\beta\sqrt{\tau}}{\alpha} \cdot H_n^{(\alpha)}(x+\sigma\beta\sqrt{\tau}) - H_n^{(\alpha)\prime}(x+\sigma\beta\sqrt{\tau}) \right) \cdot z^{(\beta)}(x)dx \\
&= -\frac{\beta}{\alpha} \int_{-d_2-\sigma\beta\sqrt{\tau}}^{\infty} H_n^{(\alpha)}(x+\sigma\beta\sqrt{\tau}) \cdot z^{(\beta)\prime}(x)dx \\
&\quad + \int_{-d_2-\sigma\beta\sqrt{\tau}}^{\infty} \left[\frac{\sigma\beta\sqrt{\tau}}{\alpha} \cdot H_n^{(\alpha)}(x+\sigma\beta\sqrt{\tau}) - H_n^{(\alpha)\prime}(x+\sigma\beta\sqrt{\tau}) \right] \cdot z^{(\beta)}(x)dx \\
&= \frac{\beta}{\alpha} \cdot H_n^{(\alpha)}(-d_2) \cdot z^{(\beta)}(-d_2-\sigma\beta\sqrt{\tau}) + \frac{\sigma\beta\sqrt{\tau}}{\alpha} \cdot I_n^{(\alpha,\beta)} + \frac{\beta-\alpha}{\alpha} \cdot \frac{n}{\alpha} \cdot I_{n-1}^{(\alpha,\beta)}.
\end{aligned}$$

We proceed in a similar way to compute the second term of the option price as

$$e^{-r\tau} \int_{-d_2}^{\infty} K \cdot \widetilde{p}_\tau(y)dy = e^{-r\tau} \cdot K \cdot \sum_{n=0}^{\infty} a_n^{(\alpha,\beta)} \cdot J_n^{(\alpha,\beta)},$$

where

$$J_n^{(\alpha,\beta)} := \int_{-d_2}^{\infty} H_n^{(\alpha)}(x) \cdot z^{(\beta)}(x)dx.$$

To establish the recurrence relation for the sequence $J_n^{(\alpha,\beta)}$, we use again the properties (2.1), (2.4) and integration by parts to obtain:

$$\begin{aligned}
J_{n+1}^{(\alpha,\beta)} &= \int_{-d_2}^{\infty} \left(\frac{x}{\alpha} \cdot H_n^{(\alpha)}(x) - H_n^{(\alpha)\prime}(x) \right) \cdot z^{(\beta)}(x)dx \\
&= \frac{\beta}{\alpha} \cdot H_n^{(\alpha)}(-d_2) \cdot z^{(\beta)}(-d_2) + \frac{\beta-\alpha}{\alpha} \cdot \frac{n}{\alpha} \cdot J_{n-1}^{(\alpha,\beta)}.
\end{aligned}$$

Getting explicit expressions for $I_0^{(\alpha,\beta)}$, $I_1^{(\alpha,\beta)}$ and $J_0^{(\alpha,\beta)}$, $J_1^{(\alpha,\beta)}$ completes the proof. □

Proof of Proposition 3.1: Similarly to the proof of Theorem 3.1, we have:

$$
\begin{aligned}
S_t \cdot e^{-q\tau} \cdot \quad |\Pi_1^{(\alpha,\beta)} - \Pi_{1,M}^{(\alpha,\beta)}| &= e^{-r\tau} \left| \int_{-d_2}^{\infty} S_t \cdot e^{\mu\tau + \sigma\sqrt{\tau}x} \cdot \left[\widetilde{p}_\tau(x)dx - z^{(\beta)}(x) R_M^{(\alpha,\beta)}(x)\right] dx \right| \\
&\leq e^{-r\tau} \cdot \int_{-d_2}^{\infty} S_t \cdot e^{\mu\tau + \sigma\sqrt{\tau}x} \cdot \left|\widetilde{p}_\tau(x) - z^{(\beta)}(x) R_M^{(\alpha,\beta)}(x)\right| dx \\
&\leq S_t \cdot e^{(\mu - r)\tau} \cdot \left|\left| \frac{\widetilde{p}_\tau(x)}{z^{(\beta)}(x)} - R_M^{(\alpha,\beta)}(x) \right|\right|_{\mathcal{L}^2_\alpha} \cdot \eta^{(\alpha,\beta)}(\sigma)
\end{aligned}
$$

with

$$
\begin{aligned}
[\eta^{(\alpha,\beta)}(\sigma)]^2 :=& \int_{-d_2}^{\infty} \left[e^{\sigma\sqrt{\tau}x} \frac{z^{(\beta)}(x)}{w^{(\alpha)}(x)} \right]^2 w^{(\alpha)}(x) dx \\
=& \sqrt{\frac{\alpha}{2\pi\beta(2\alpha - \beta)}} \cdot \exp\left\{ \frac{4\alpha\beta\sigma^2\tau}{2\alpha - \beta} \right\} \cdot N\left(d_2 \sqrt{\frac{2\alpha - \beta}{\alpha\beta}} + 2\sigma\sqrt{\frac{\tau\alpha\beta}{2\alpha - \beta}} \right).
\end{aligned}
$$

Analogously,

$$
|\Pi_2^{(\alpha,\beta)} - \Pi_{2,M}^{(\alpha,\beta)}| \leq \left|\left| \frac{\widetilde{p}_\tau(x)}{z^{(\beta)}(x)} - R_M^{(\alpha,\beta)}(x) \right|\right|_{\mathcal{L}^2_\alpha} \cdot \eta^{(\alpha,\beta)}(0).
$$

From the orthogonality property (2.4) of Generalized Hermite polynomials it follows that

$$
\begin{aligned}
\left|\left| \frac{\widetilde{p}_\tau(x)}{z^{(\beta)}(x)} - R_M^{(\alpha,\beta)}(x) \right|\right|^2_{\mathcal{L}^2_\alpha} &= \int_{\mathbb{R}} \left| \sum_{n=M+1}^{\infty} a_{\tau,n}^{(\alpha,\beta)} H_n^{(\alpha)}(x) \right|^2 w^{(\alpha)}(x) dx \\
&= \sum_{n=M+1}^{\infty} \left[a_{\tau,n}^{(\alpha,\beta)}\right]^2 \int_{\mathbb{R}} \left[H_n^{(\alpha)}(x)\right]^2 w^{(\alpha)}(x) dx \\
&= \sqrt{2\pi\alpha} \sum_{n=M+1}^{\infty} \left[a_{\tau,n}^{(\alpha,\beta)}\right]^2 \cdot n! \cdot \alpha^{-n}. \qquad \text{(A.4)}
\end{aligned}
$$

For the expansion coefficients $a_{\tau,n}^{(\alpha,\beta)}$ from formula (2.6) we obtain:

$$
\begin{aligned}
|a_{\tau,n}^{(\alpha,\beta)}| &= \frac{1}{\sqrt{2\pi\alpha}} \frac{\alpha^n}{n!} \cdot \left| \int_{\mathbb{R}} \frac{\widetilde{p}_\tau(x)}{z^{(\beta)}(x)} \cdot H_n^{(\alpha)}(x) \cdot e^{-\frac{x^2}{2\alpha}} dx \right| \\
&= \frac{\alpha^{n-\frac{1}{2}}}{\sqrt{2\pi} \cdot n!} \cdot \left| \int_{\mathbb{R}} \frac{\widetilde{p}_\tau(x)}{z^{(\beta)}(x)} \cdot d\left(e^{-\frac{x^2}{2\alpha}}\right)^{(n-1)} \right| \\
&= \frac{\alpha^{n-\frac{1}{2}}}{\sqrt{2\pi} \cdot n!} \cdot \left| \int_{\mathbb{R}} \frac{\partial}{\partial x} \frac{\widetilde{p}_\tau(x)}{z^{(\beta)}(x)} \cdot \left(e^{-\frac{x^2}{2\alpha}}\right)^{(n-1)} dx \right| \\
&= \frac{\alpha^{n-\frac{1}{2}}}{\sqrt{2\pi} \cdot n!} \cdot \left| \int_{\mathbb{R}} \frac{\partial^k}{\partial x^k} \frac{\widetilde{p}_\tau(x)}{z^{(\beta)}(x)} \cdot \left(e^{-\frac{x^2}{2\alpha}}\right)^{(n-k)} dx \right| \\
&= \frac{\alpha^{n-\frac{1}{2}}}{\sqrt{2\pi} \cdot n!} \cdot \left| \int_{\mathbb{R}} \frac{\partial^k}{\partial x^k} \frac{\widetilde{p}_\tau(x)}{z^{(\beta)}(x)} \cdot H_{n-k}^{(\alpha)}(x) \cdot e^{-\frac{x^2}{2\alpha}} dx \right|,
\end{aligned}
$$

where we used (2.2) and (2.5) to get $\frac{\widetilde{p}_\tau(x)}{z^{(\beta)}(x)}\big(e^{-\frac{x^2}{2\alpha}}\big)^{(l)}\Big|_{\pm\infty} = 0$ for $l = 1, 2, \dots$. Further, by the Cauchy-Schwarz inequality we have:

$$
\begin{aligned}
|a^{(\alpha,\beta)}_{\tau,n}| &\leq \frac{\alpha^{n-\frac{1}{2}}}{\sqrt{2\pi}\cdot n!}\cdot \big|\big|H^{(\alpha)}_{n-k}(x)\big|\big|_{\mathcal{L}^2_\alpha}\cdot \left|\left|\frac{\partial^k}{\partial x^k}\frac{\widetilde{p}_\tau(x)}{z^{(\beta)}(x)}\right|\right|_{\mathcal{L}^2_\alpha} \\
&\leq \frac{\alpha^{\frac{n+k}{2}-\frac{1}{4}}}{(2\pi)^{1/4}}\cdot\frac{\sqrt{(n-k)!}}{n!}\cdot \left|\left|\frac{\partial^k}{\partial x^k}\frac{\widetilde{p}_\tau(x)}{z^{(\beta)}(x)}\right|\right|_{\mathcal{L}^2_\alpha}.
\end{aligned}
$$

Inserting the latter into (A.4) yields:

$$
\left|\left|\frac{\widetilde{p}_\tau(x)}{z^{(\beta)}(x)} - R^{(\alpha,\beta)}_M(x)\right|\right|^2_{\mathcal{L}^2_\alpha} \leq \alpha^k\cdot \left|\left|\frac{\partial^k}{\partial x^k}\frac{\widetilde{p}_\tau(x)}{z^{(\beta)}(x)}\right|\right|^2_{\mathcal{L}^2_\alpha}\cdot \sum_{n=M+1}^{\infty}\frac{(n-k)!}{n!},
$$

where

$$
\begin{aligned}
\sum_{n=M+1}^{\infty}\frac{(n-k)!}{n!} &= \sum_{n=M+1}^{\infty}\frac{1}{(n-(k-1))\cdot\ldots\cdot n} \\
&= \frac{1}{k-1}\sum_{n=M+1}^{\infty}\left[\frac{1}{(n-(k-1))\cdot\ldots\cdot(n-1)} - \frac{1}{(n-(k-2))\cdot\ldots\cdot n}\right] \\
&= \frac{1}{k-1}\cdot\frac{1}{(M-(k-2))\cdot\ldots\cdot M}
\end{aligned}
$$

with the series converging for any $k \geq 2$. Accounting for

$$
\begin{aligned}
&|C\big(K,\tau,S_t,r,q,\mu,\sigma,a^{(\alpha,\beta)}_n\big) - C_M\big(K,\tau,S_t,r,q,\mu,\sigma,a^{(\alpha,\beta)}_n\big)| \\
&\qquad\qquad \leq S_t\cdot e^{-q\tau}\cdot|\Pi^{(\alpha,\beta)}_1 - \Pi^{(\alpha,\beta)}_{1,M}| + K\cdot e^{-r\tau}\cdot|\Pi^{(\alpha,\beta)}_2 - \Pi^{(\alpha,\beta)}_{2,M}|
\end{aligned}
$$

and rearranging the terms completes the proof. □

A.3 Calibration of expansion coefficients

Proof of Proposition 3.2: Denote by $\rho_\tau(\cdot)$ the risk-neutral density of $S_{t+\tau}$. Making the change of variable $S = S_t\cdot e^x$ in (2.6), we obtain:

$$
\begin{aligned}
a^{(\alpha,\beta)}_{\tau,n} &= \frac{\sqrt{2\pi}\cdot\alpha^{n-\frac{1}{2}}}{n!}\int_0^\infty \rho_\tau(S)H^{(\alpha)}_n\left(-\frac{\log(S_t/S)+\mu\tau}{\sigma\sqrt{\tau}}\right)\widetilde{z}^{(\alpha,\beta)}\left(\frac{\log(S_t/S)+\mu\tau}{\sigma\sqrt{\tau}}\right)dS \\
&= \frac{\sqrt{2\pi}\cdot\alpha^{n-\frac{1}{2}}}{n!}\cdot\mathbb{E}\big[\widetilde{z}^{(\alpha,\beta)}(d_2(S_{t+\tau}))\cdot H^{(\alpha)}_n(-d_2(S_{t+\tau}))\big],
\end{aligned}
$$

where $\mathbb{E}$ is the expectation under the measure ρ_τ. Excluding the constant scaling factor in front, we recognize $a^{(\alpha,\beta)}_{\tau,n}$ as the (undiscounted) value at time t of the payoff

$$
H(S_{t+\tau}) = \widetilde{z}^{(\alpha,\beta)}(d_2(S_{t+\tau}))\cdot H^{(\alpha)}_n\left(-d_2(S_{t+\tau})\right),
$$

received at time $t+\tau$. Following Bakshi et al. (2003), we now expand the payoff in a continuum of OTM European vanilla call and put options using as a cutoff point the forward price $F_t = S_t \cdot e^{(r-q)\tau}$ to obtain:

$$\begin{aligned}
\widetilde{z}^{(\alpha,\beta)}(d_2(S_{t+\tau})) \cdot H_n^{(\alpha)}(-d_2(S_{t+\tau})) &= \left[\widetilde{z}^{(\alpha,\beta)}(d_2(F_t)) \cdot H_n^{(\alpha)}(-d_2(F_t)) - H'_S(F_t) \cdot F_t\right] \\
&\quad + H'_S(F_t) \cdot S_{t+\tau} + \int_0^{F_t} H''_{SS}(K) \cdot (K - S_{t+\tau})^+ dK \\
&\quad + \int_{F_t}^{\infty} H''_{SS}(K) \cdot (S_{t+\tau} - K)^+ dK,
\end{aligned}$$

or after taking the expectation of both left and right parts:

$$\begin{aligned}
\mathbb{E}\left[\widetilde{z}^{(\alpha,\beta)}(d_2(S_{t+\tau})) \cdot H_n^{(\alpha)}(-d_2(S_{t+\tau}))\right] &= \widetilde{z}^{(\alpha,\beta)}(d_2(F_t)) \cdot H_n^{(\alpha)}(-d_2(F_t)) \\
&\quad + e^{r\tau} \cdot \int_0^{F_t} H''_{SS}(F_t) \cdot P(K,\tau,S_t) dK \\
&\quad + e^{r\tau} \int_{F_t}^{\infty} H''_{SS}(K) \cdot C(K,\tau,S_t) dK,
\end{aligned}$$

where we used that $P(K,\tau,S_t) = e^{-r\tau} \cdot \mathbb{E}_t(K - S_{t+\tau})^+$ and $C(K,\tau,S_t) = e^{-r\tau} \cdot \mathbb{E}_t(S_{t+\tau} - K)^+$. Finally, we use that

$$\begin{aligned}
H''_{SS}(K) = \frac{\widetilde{z}^{(\alpha,\beta)}(d_2(K))}{K^2\sigma^2\tau} \cdot \frac{\beta-\alpha}{\alpha\beta} \cdot \Bigg\{ \left[d_2^2(K) \frac{\beta-\alpha}{\alpha\beta} - \sigma\sqrt{\tau} d_2(K) - 1 \right] H_n^{(\alpha)}(-d_2(K)) \\
+ \frac{n}{\alpha} \cdot \left[2 \cdot \frac{\beta-\alpha}{\alpha\beta} \cdot d_2(K) - \sigma\sqrt{\tau} \right] H_{n-1}^{(\alpha)}(-d_2(K)) + \frac{n(n-1)}{\alpha^2} H_{n-2}^{(\alpha)}(-d_2(K)) \Bigg\}
\end{aligned}$$

to complete the proof. □

B Empirical study for 2007-2008

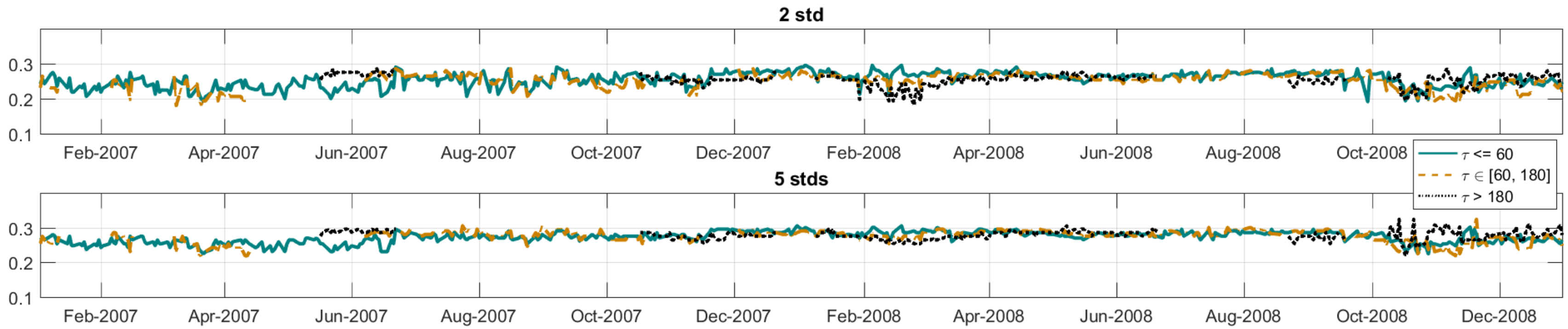
Table B.1: MSE of the Gauss-Hermite expansion in 2007-2008.

Days-to-exp.\ M	10	12	14	16	18	20	22	24	26	28	30	32	34	36	38	40
≤ 60	4.32	2.18	2.00	1.56	1.46	1.34	1.27	1.22	1.18	1.15	1.13	1.11	1.09	1.08	1.06	1.06
$(60, 180]$	11.31	5.26	5.14	3.96	1.49	1.21	1.15	1.09	1.02	1.06	1.04	1.06	1.04	1.03	1.01	1.03
> 180	43.86	15.36	24.84	12.44	20.95	13.02	7.99	3.39	3.51	2.74	2.78	2.85	2.66	2.52	2.62	2.47
all	48.12	17.12	26.70	13.63	21.52	13.02	7.55	3.27	3.11	2.39	2.24	2.18	1.83	1.52	1.35	1.21

NOTES: The reported numbers are the average mean square errors (MSE) between model implied and quoted (bid-ask average) option prices relative to the average option price in a given maturity class, in percent. Gauss-Hermite expansion ($\alpha = 0.5$, $\beta = 1$) truncated after $M = 10 - 40$ terms, FIT method. The averages are computed for the sample period from January 1st, 2007 to December 31st, 2008 and across option maturities $\tau < 60$, $\tau \in (60, 180]$ and $\tau > 180$ days to expiration.

Figure B.1: Tail parameter (4.2) in 2007-2008, FIT method.

(a) $M = 10 - 40$



(b) $M = 20 - 40$

NOTES: Time series of tail parameters (4.2) for the sample period from January 1st, 2007 to December 31st, 2008, FIT method. The plotted values are the average tail parameters across option maturities $\tau < 60$, $\tau \in (60, 180]$ and $\tau > 180$ days to expiration. Tail parameters are estimated based on the empirical $\mathcal{L}^2_\alpha$-convergence of approximating polynomials benchmarked by 2 and 5 average standard deviations of Gauss-Hermite exponential growths (across all observation dates and option maturities) in excess to the Gauss-Hermite exponential growth for a given observation date and option maturity. $M = 10 - 40$ (a) and $M = 20 - 40$ (b) refer to the range of approximating polynomials employed in computing exponential growths of $\mathcal{L}^2_\alpha$-norms.

Table B.2: Exponential growths of $\mathcal{L}^2$-norms of Gauss-Hermite approximating polynomials in 2007-2008.

Days-to-exp.	≤ 60			$(60, 180]$			> 180			all		
M	10-30	10-40	20-40	10-30	10-40	20-40	10-30	10-40	20-40	10-30	10-40	20-40
av. gr., 10^{-3}	2.71	1.95	1.19	2.83	1.82	0.86	3.12	2.37	1.42	2.90	1.99	1.13
av. std, 10^{-3}	0.49	0.32	0.10	0.62	0.40	0.12	0.88	0.60	0.27	0.68	0.40	0.13

NOTES: The reported numbers are the average exponential growths and their average Newey-West standard errors estimated from the linear model (4.5) with $\alpha = 0.5$ and $M = 10 - 30$, $10 - 40$ and $20 - 40$. The averages are computed for the sample period from January 1st, 2007 to December 31st, 2008 and across option maturities $\tau < 60$, $\tau \in (60, 180]$ and $\tau > 180$ days to expiration.

Table B.3: Sample properties of S&P 500 index options in 2007-2008.

		Moneyness S/K	Days-to-expiration			Subtotal
			≤ 60	$(60, 180]$	> 180	
Calls	OTM	≤ 0.94	2.68 (1.00) [716.98] {24886}	7.96 (1.65) [349.22] {19404}	34.26 (3.26) [143.38] {29391}	[391.32] {73681}
		$(0.94, 0.97]$	9.44 (1.38) [2539.70] {7656}	32.06 (2.57) [872.63] {3900}	102.94 (3.83) [276.31] {3809}	[1555.46] {15365}
	ATM	$(0.97, 1.00]$	21.23 (1.93) [4053.17] {8075}	49.89 (2.71) [1818.41] {4239}	124.14 (3.69) [379.27] {3777}	[2602.08] {16091}
		$(1.00, 1.03]$	44.30 (2.49) [1891.51] {7510}	73.80 (2.75) [867.42] {4065}	147.75 (3.77) [234.97] {3755}	[1214.19] {15330}
	ITM	$(1.03, 1.06]$	74.37 (2.62) [428.70] {6481}	101.15 (2.84) [138.12] {3595}	169.55 (3.65) [51.66] {3498}	[254.58] {13574}
		> 1.06	219.55 (2.85) [64.73] {30095}	266.06 (2.96) [25.51] {17431}	377.53 (3.85) [8.27] {30483}	[33.90] {78009}
Puts	ITM	≤ 0.94	236.56 (3.48) [103.85] {23577}	271.54 (3.54) [76.45] {17638}	369.43 (4.51) [24.28] {28747}	[64.25] {69962}
		$(0.94, 0.97]$	70.93 (2.72) [547.81] {7143}	87.58 (2.86) [382.55] {3887}	132.63 (3.82) [186.87] {3809}	[411.87] {14839}
	ATM	$(0.97, 1.00]$	39.89 (2.48) [2557.59] {7904}	63.31 (2.74) [1867.79] {4240}	110.35 (3.67) [492.64] {3775}	[1884.18] {15919}
		$(1.00, 1.03]$	22.12 (1.97) [5252.61] {7591}	47.51 (2.65) [2079.04] {4065}	94.62 (3.66) [722.92] {3754}	[3311.99] {15410}
	OTM	$(1.03, 1.06]$	13.49 (1.64) [3735.75] {6742}	36.65 (2.63) [1477.50] {3595}	79.74 (3.52) [515.12] {3499}	[2334.52] {13836}
		> 1.06	4.37 (1.03) [1649.84] {36984}	12.56 (1.68) [803.01] {20183}	33.02 (2.92) [253.61] {33322}	[946.80] {90489}
Subtotal			{174644}	{106242}	{151619}	{432505}

NOTES: The reported numbers are the average \$ quoted bid-ask mid point prices, the average \$ bid-ask spreads (best bid minus best ask price) in parenthesis, the average trading volumes per contract per day in square brackets and the total number of observations in curly brackets. The sample period is from January 1st, 2007 to December 31st, 2008 with total of $432,505$ options. S denotes the spot S&P 500 index level and K is the strike price. OTM, ATM and ITM stand for out-of-the-money, at-the-money and in-the-money options respectively.

Chapter III

Testing the Stochastic Disorder Model on Stock Markets

Abstract. This paper examines real-time applications of quickest disorder detection techniques to stock market timing. The focus is on the stochastic disorder model by Shiryaev, Zhitlukhin, and Ziemba (2014, 2015), Zhitlukhin and Ziemba (2016) and their optimal stopping rule. The model uses sequential price data to identify a directional change in the market trend and determines the optimal exit moment from a long position in a bubble-like market. Together with the sensitivity analysis of the exit rule's signals with respect to model parameters, we study out-of-sample performance of the entry-exit investment strategy that exploits signals from the rule. Using daily historical data on the S&P 500, we find that the entry-exit strategy underperforms the buy-and-hold strategy over the whole testing period 1965–2016, but outperforms it in the fall of 1987 and during the bear market of 2007–2009.

Keywords: Bubbles; market timing; quickest disorder detection.
JEL classification: C53; G17.

1 Introduction

This paper investigates real-time applications of the optimal stopping rule (hereafter, exit rule) by Shiryaev and Zhitlukhin (2012a), Shiryaev, Zhitlukhin, and Ziemba (2014, 2015) and Zhitlukhin and Ziemba (2016) to stock market timing. We give insights into the economics of the exit rule and conduct a detailed sensitivity analysis of its signals with respect to model parameters. We then employ the results to develop a fully-fledged investment strategy that exploits signals from the exit rule to enter and exit the market. Out-of-sample performance of the strategy is tested on the historical data on daily prices of the S&P 500 index in 1965–2016.

The innovative approach of Shiryaev et al. (2014, 2015), Zhitlukhin and Ziemba (2016) consists in utilization of the mathematical theory of quickest disorder detection for the problem of a timely exit from a bubble-like market.[1] Their notion of the bubble-like market refers exclusively to the behaviour of a price process, namely to presence of directional changes in the price trend. In this respect the methodology differs from the classical literature that focuses on economic reasons of the bubble-like price behaviour and makes predictions by testing for presence of rational and behavioural bubbles, e.g., Johansen, Sornette, and Ledoit (1999), Li and Xue (2009), Phillips, Wu, and Yu (2011), Wu (1997). The approach of Shiryaev, Zhitlukhin, and Ziemba (2014, 2015), Zhitlukhin and Ziemba (2016), in turn, can be applied to any market, as perception of market trends is subjective, and almost any price pattern can be considered as experiencing the bubble-like behaviour.

The exit rule determines the optimal moment to liquidate the underlying based exclusively on the price process, while sequentially observing it. It requires forming a prior opinion on current and future market regimes (drift and volatility), selecting an observation window and assessing the probability of a regime change within given observation window. The mathematical foundations of the exit rule are given in Shiryaev and Zhitlukhin (2012a,b) and Zhitlukhin (2014).

Empirical tests in Shiryaev et al. (2014, 2015), Zhitlukhin and Ziemba (2016) revealed potential of the exit rule for market timing and risk management purposes. The authors applied the rule on several historical crises and crashes and found that in the majority of cases the exit rule could capture about 70-80% of the maximum (pre-crash) return.[2]

This paper extends the previous empirical studies by running out-of-sample tests of the exit rule's performance. Unlike the original works, we apply the exit rule to multiple market scenarios, also when no crash/directional change occurs within

[1] The history of quickest disorder detection methods goes back to pioneering works of Shewhart in 1920s, and the first results by Page, Roberts, Shiryaev and Stewhart in 1950-1960s, see Shiryaev (2010) for an overview. The theory found its successful application in production quality control, radiolocation and information security, but only recently has drawn attention of financial literature, see, e.g., Shiryaev (2002) for an overview of major results.

[2] See Ziemba, Lleo, and Zhitlukhin (2017) for an overview of the results.

the observation period. The dynamic structure of the entry-exit strategy allows to overcome sample as well as look-ahead biases of past studies. Our in-sample sensitivity tests, on the other hand, assess the value of information on the present and the future of the underlying for performance of the exit rule in market timing. The study therefore provides a different perspective on real-time applications of the exit rule, when the future price pattern is known neither quantitatively nor qualitatively.

We find that the exit rule's signals vary significantly with the forward-looking observation window as well as with the certainty of belief that a change in the price trend would occur within the given observation window. The pattern is particularly pronounced for the observation window. Thus, the closer is the latest observation date to the actual change of the price trend, the shorter is the delay of the signal and the higher is the realized return. The result is intuitive as the observation window together with the probability of a structural break within this time horizon provide quantitative information regarding the actual moment of the upcoming change. We also find that mistakes in estimation of current and future market regimes, that correspond to qualitative information regarding the change, have significantly less effect on the signal.

Our second line of research provides a methodology to calibrate model parameters of the exit rule in order to apply it in a dynamic manner, so that one can benefit from price appreciation and depreciation. By using the same exit rule in a repetitive manner we also account for optimality of the exit rule *on average* only, as opposed to optimality on a given price realization, which has been analysed in the previous works.

We find that the entry-exit strategy based on the exit rule generates positive returns in-sample, but most of these returns disappear out-of-sample. The law Sharpe ratio of the strategy indicates high volatility of its returns. The entry-exit strategy cannot outperform the buy-and-hold strategy over the whole testing period from 1965 to 2016 either in- or out-sample. But the strategy outperforms the buy-and-hold rule during the major bear markets in the fall of 1987 and during 2007–2009.

Our out-of-sample tests suggest that the exit rule reacts in a timely manner to sharp price declines such as the Crash of 1987, which complies with findings in Zhitlukhin and Ziemba (2016). In normal markets however delays in the exit rule's signals are so long that most of returns, that were realized during a favourable market trend, get lost because the position remains unadjusted during the subsequent unfavourable market period until the rule signals to exit. Moreover, if the strategy misidentifies the prevailing market regime, it takes long until it can coordinate with the actual market regime.

Based on our empirical results we conclude that the exit rule cannot be used as a single investment instrument. Its ability to identify significant market corrections, on the other hand, suggests its potential applicability for insurance purposes.

Related literature on quickest disorder detection techniques in the context of stock market timing includes papers by, e.g., Gapeev (2010), Nguyen, Tie, and Zhang (2014), Sokko (2015), Zhitlukhin and Shiryaev (2013). Zhitlukhin and Shiryaev (2013), Sokko (2015) consider alternative penalty functions, respectively, linear and exponential penalty functions as well as that minimizing the expected delay of the signal. Gapeev (2010) derives an optimal stopping rule for a multiple disorder problem. Nguyen et al. (2014) study optimal trading rules under a switchable mean-reversion model. This paper contributes to the literature by exploring practical rather than theoretical applications of the methodology.

Technical trading and prediction of future price dynamics based exclusively on the information from observed market prices have been also addressed in Glabadanidis (2014, 2015). The studies document the market timing abilities of the moving average (MA) strategy and show that, unlike the entry-exit strategy of this paper, the MA strategy outperforms the buy-and-hold strategy. Jiang, Zhou, Sornette, Woodard, Bastiaensen, and Cauwels (2010), Zhang, Sornette, Balcilar, Gupta, Ozdemir, and Yetkiner (2016), in turn, focus on the informative content of both market trend and volatility. They show that superexponential price behaviour together with increased market volatility correspond to presence of a bubble, hence imply an upcoming crash. Similar to findings of this paper, their approach to predict future market corrections however does not produce a stable estimate of the actual moment of the bubble burst (Forró, 2015).

This paper is structured as follows. Section 2 looks into the economics of the exit rule and illustrates the influence of model parameters on its signals. Section 3 conducts the sensitivity analysis. Section 4 proceeds with out-of-sample tests of the entry-exit strategy. The final section summarizes the main results and conclusions.

2 Stock market model and the optimal exit rule by Zhitlukhin and Ziemba (2016)

In this section we make a short summary and provide an economic interpretation of the stochastic disorder model of a stock market by Shiryaev et al. (2014, 2015), Zhitlukhin and Ziemba (2016). The exit rule arises as the solution of the expected utility maximization problem of a trader who holds a subjective opinion on the market.

Agent's opinion on the market. An agent (trader) thinks of the market in terms of *risk factors*.[3] He observes each risk factor as a sequence of prices S_t at

[3] A risk factor may be an individual asset, a basket of assets or the overall market.

discrete times $t = \ldots, -1, 0, 1, \ldots$.

At time $t = 0$ the trader selects a certain risk factor and forms an opinion on it.[4] The opinion contains a view on the current regime of the factor as well as a forward looking statement. It sounds as follows: *"Currently the market experiences 4% annual trend and 20% annual volatility. Within 5 years, with 75 % probability, the trend will change to -5% and the volatility to 25%"* (here the risk factor is the overall market). The opinion contains:

(a) a view on the current regime of the factor (current trend $\mu_1 = 4\%$ and volatility $\sigma_1 = 20\%$),[5]

(b) a view on the future regime of the factor (future trend $\mu_2 = -5\%$ and volatility $\sigma_2 > 0 = 25\%$),

(c) forward looking time horizon T (which corresponds to the notion of "future", 5 years),

(d) certainty of beliefs p (75%).

The trader is certain about both current (μ_1, σ_1) and future (μ_2, σ_2) regimes of the risk factor, but is uncertain whether the change of regime will happen within horizon T and if so, when.

Under assumption of independent and normally distributed one-period logarithmic returns, the trader's opinion on the risk factor may be formally stated as follows:

$$X_t := \log \frac{S_t}{S_{t-1}} = \begin{cases} \mu_1 + \sigma_1 \cdot \xi_t, & t < \theta, \\ \mu_2 + \sigma_2 \cdot \xi_t, & t \geq \theta \end{cases} \tag{2.1}$$

with ξ_t independent standard normally distributed random variables. The moment θ of change of regime (disorder moment) is *unobservable* by the trader. It is independent of ξ_t, $t = 0, 1, \ldots$, and occurs between $t = 1$ and $t = T$ with probability p, equally likely for each $t \in [1, T]$, i.e.,

$$\begin{aligned} &\mathbb{P}(\theta = t \mid \theta \leq T) = 1/T \text{ for each } t = 1, \ldots, T, \\ &\mathbb{P}(\theta \leq T) = p. \end{aligned} \tag{2.2}$$

Agent's decision problem. At time $t = 0$, based on the subjective opinion $(\mu_1, \sigma_1, \mu_2, \sigma_2, T, p)$, the trader decides to invest amount w_0 in the risk factor S. The intention is to keep the position (in full) until future time T (determined by the forward looking horizon of the agent's opinion on the risk factor) or liquidate it (to the full) at some $\tau < T$.

[4] The choice of the risk factor is itself an opinion of the trader.

[5] As the current trend (and, to a lesser extent, volatility) is difficult to measure, it is considered an opinion. E.g., if the trend is estimated as mean of past one-period logarithmic returns, the value varies significantly with the range of past data employed in estimation.

The decision to liquidate (exit) at time $\tau = t'$ with $t' \in [0, T]$ depends on realized market prices $(S_l)_{l \le t'}$ up to time t' ($\tau \in \mathcal{M}(S)$).[6] The optimality criteron is to maximize the expected utility from terminal wealth w_τ:

$$\tau_\eta^* = \arg \max_{\tau \in \mathcal{M}(S)} \mathbb{E}_0\, U^\eta(w_\tau) \tag{2.3}$$

with

$$w_\tau = \begin{cases} w_0 \cdot (S_\tau / S_0), & \text{in case of a long position } (\mu_1 > 0), \\ w_0 \cdot (1 - (1 - S_0/S_\tau)/\mathcal{M}), & \text{in case of a short position } (\mu_1 < 0) \end{cases}$$

($\mathcal{M}$ stands for margin requirement), $\mathbb{E}_0$ the expected value at time 0 given dynamics (2.1)–(2.2) and CRRA instantaneous utility $U^\eta(\cdot)$ with risk aversion η,

$$U^\eta(w) = \begin{cases} \frac{1}{1-\eta} w^{1-\eta}, & \eta \in [0,1) \cup (1, \infty), \\ \ln(w), & \eta = 1. \end{cases}$$

$\tau_\eta^* = 0$ means that the trader makes no investment in the risk factor, whereas $\tau_\eta^* = T$ implies that the position is held until latest observation time T.

Optimal exit rule from a long position. When $\mu_1 > 0$, the problem (2.3) becomes

$$\tau_\eta^* = \arg \max_{\tau \in \mathcal{M}(S)} \mathbb{E}_0\, U^\eta(S_\tau). \tag{2.4}$$

Zhitlukhin (2014) finds the solution in the form of

$$\tau_\eta^* = \inf\{0 \le t \le T \mid \pi_t \ge b_\eta^*(t)\} \tag{2.5}$$

with

$$\pi_t = \mathbb{P}(\theta \le t \mid S_0, \dots, S_t) \tag{2.6}$$

the posterior probability that a change of regime occurred by t, and $b_\eta^*(t) = b_\eta^*(\mu_1, \sigma_1, \mu_2, \sigma_2, T, p;\ t)$ a decreasing time-dependent threshold (see Appendix A for details).[7]

The exit rule (2.5) determines the optimal moment to exit from a long position in the risk factor as the first moment when the posterior probability π_t hits the time-dependent threshold $b_\eta^*(t)$. The threshold function b_η^* is computed at time $t = 0$ and kept until the end of the observation period ($t = T$). The posterior probability π_t is recomputed sequentially with help of Bayesian updating using new price realizations.

The time-dependent structure of the threshold $b_\eta^*(t)$ accounts for unconditional

[6] Formally, τ is a stopping time with respect to the price sequence S_t.

[7] In two special cases the problem (2.4) is degenerate. Namely, when $2\mu_1 < -\sigma_1^2 \cdot (1 - \eta)$, the trader's expected utility decreases with holding time independent of presence of regime change, therefore $\tau_\eta^* \equiv 0$. Analogously, when $2\mu_1 \ge -\sigma_1^2 \cdot (1 - \eta)$ and $2\mu_2 \ge -\sigma_2^2 \cdot (1 - \eta)$, the trader's expected utility increases with holding time, hence $\tau_\eta^* \equiv T$.

probability of regime change to be increasing in t. The latter also implies that the exit rule becomes more sensitive to drops in the risk factor's price towards the end of the observation period (i.e., $t = T$). A similar price drop at the beginning of the observation period would be often taken for market volatility rather than a change of regime.

Optimal exit rule from a short position. When $\mu_1 < 0$ and $\eta = 0$ (i.e., the trader maximizers the expected return on investment), the problem (2.3) is identical to (2.4) with $\widetilde{\mu}_1 = -\mu_1$ and $\widetilde{\mu}_2 = -\mu_2$. Hence it admits solution (2.5) with the corresponding adjustment in market regimes. No solution is known for the general case $\eta \neq 0$.

2.1 Application of the exit rule to market timing

The first application of the exit rule to market timing is due to Shiryaev et al. (2014, 2015), Zhitlukhin and Ziemba (2016). The authors tested the exit rule on historical price data during major bubbles and crashes, such as the Great Crash in the DJIA in 1929, the 1987's crash of S&P 500, the internet bubble crash of 2000–2002, the AAPL bubble in 2009–2012. They found that in the majority of cases the rule could provide a timely signal to liquidate the position soon after the bubble burst, and closing the position on this date would capture about 70–80% of the maximum pre-crash return.

2.1.1 Fragility of the exit rule's signals

The main difficulty related to real-time applications of the exit rule (2.5) consists in selecting values of 7 model parameters, namely (μ_1, σ_1), (μ_2, σ_2), (T, p) and η, which are necessary to set-up the rule. Table 2.1 lays out dependence of the threshold function and the posterior probability on model parameters. Thus, all 6 market-related variables, (μ_1, σ_1), (μ_2, σ_2), (T, p), affect both the threshold function and the posterior probability. The risk aversion η, which defines the optimality criterion, influences only the threshold function.

Figure 2.1 illustrates the effect of model parameters on signals from the exit rule (2.5) by the example of the 1987's crash in the S&P 500. The benchmark specification of the exit rule (in red) is the entry ($t = 0$) on October 1st, 1986 (red square), the latest exit ($t = T$) on December 31st, 1987, the current drift μ_1 and volatility σ_1 estimated from 1 year of past data, $\mu_2 = -\mu_1$, $\sigma_2 = \sigma_1$, $p = 75\%$ and $\eta = 3/2$.[8] Liquidating a long position in the S&P 500, formed on October 1st, 1986, on the corresponding exit date captures 90,6% of the maximum return.

When the latest exit date is shifted by 1 year forward (to December 31st, 1988), the exit rule signals only after the market has fully collapsed, realizing 66,8% of the

[8] A negative power utility with $\eta = 3/2$ has been employed in Zhitlukhin and Ziemba (2016).

Table 2.1: **Model parameters required for specification of the exit rule.**

Param.	Description	b^*_η	π_t
μ_1	Mean of 1-period log. returns prior to regime change	✓	✓
σ_1	Standard deviation of 1-period log. returns prior to regime change	✓	✓
μ_2	Mean of 1-period log. returns after regime change	✓	✓
σ_2	Standard deviation of 1-period log. returns after regime change	✓	✓
T	Forward looking observation window	✓	✓
p	Probability that regime change happens within $[1, T]$	✓	✓
η	Constant relative risk aversion	✓	✗

NOTES: Parameters required for computing the threshold function b^*_η (A.2) and updating the posterior probability π_t (A.1).

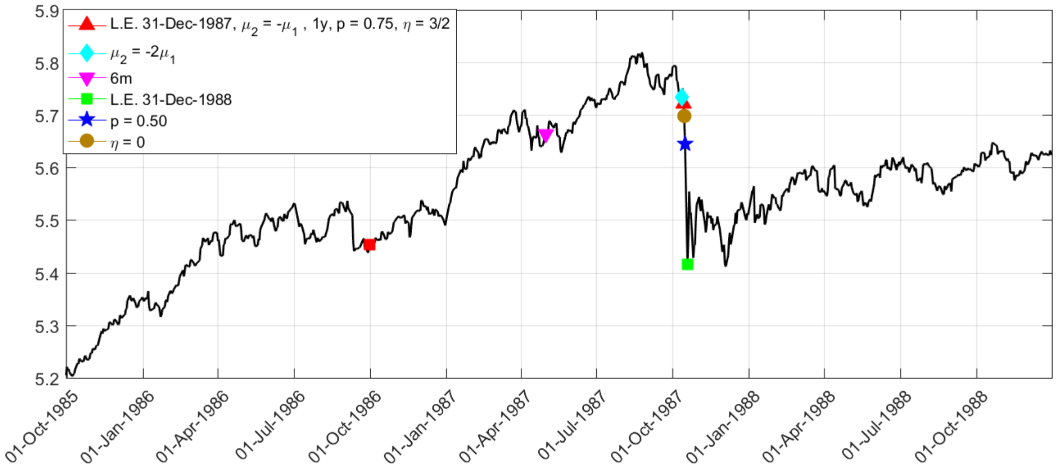
maximum return. Lowering the probability p also reduces the performance. In both cases the effect is likely due to decreased unconditional probability of regime change, hence decreased sensitivity of the exit rule to price fluctuations. On the contrary, a change $\mu_2 = -2\mu_1$ increases the performance to 91,9% of the maximum return being captured by the rule. Finally, lowering the estimation horizon for the current regime (μ_1, σ_1) to 6 months instead of 1 year results in underestimation of drift μ_1 and overestimation of volatility σ_1, with the exit rule reacting already to the price decline of summer 1987.

Our example therefore suggests that signals from the exit rule depend significantly on its specification. This dependence may be because of one or several reasons. Firstly, except for the case $\eta = 0$ (expected return maximization), "beating the market" is not an objective of the exit rule (see equation (2.3)). Secondly, the model parameters (μ_1, σ_1), (μ_2, σ_2), which refer to current and future market regimes, cannot be estimated with certainty. The observation window T and probability p are also subjective opinions of a trader. Finally, the exit rule (2.5) maximizes expected, hence average utility, which implies that the signal it produces is not necessary optimal on a given price realization. Empirical studies of the following sections investigate these issues.

3 Sensitivity analysis: Which information matters and how?

In this section we conduct a sensitivity analysis of the exit rule's signals with respect to 6 model parameters (μ_1, σ_1), (μ_2, σ_2) and (T, p), which refer to the trader's opinion on the present and the future of the risk factor. To this aim, we follow Zhitlukhin and Ziemba (2016) approach and assess the quality of the exit rule's signals by their ability to timely detect the moment of a bubble burst. We address the following questions:

Figure 2.1: Signals from the exit rule (2.5) for the 1987's crash in the S&P 500.



NOTES: The S&P 500 index in 1985–1987 and 6 signals from the exit rule (2.5) with different parameters specifications, logarithmic scale. The entry date (red square) is on October 1st, 1986, for all 6 cases. The benchmark specification (in red) is the latest exit on December 31st, 1987, estimation of regime (μ_1, σ_1) from 1 year of past data, $\mu_2 = -\mu_1$, $\sigma_2 = \sigma_1$, $p = 75\%$ and $\eta = 3/2$. 5 other signals are obtained by a) setting $\mu_2 = -2\mu_1$ (diamond), b) estimating (μ_1, σ_1) from 6 months of past data (upturned triangle), c) moving the latest exit to December 31st, 1988 (green square) d) decreasing p to 50% (star) e) changing trader's risk aversion to $\eta = 0$ (circle).

- Which information is most important?
- Is it always better to be correct or certain mistakes improve the performance?
- What is the value of information?

Similar to the study in Zhitlukhin and Ziemba (2016), we run the tests in a controlled framework with ex post identified bubble periods and ex post estimated market regimes (μ_1, σ_1) and (μ_2, σ_2). For each entry point we determine a benchmark exit point (bubble peak), to which the signal from the exit rule is then compared.

Throughout the remainder of the paper we restrict our attention to the case $\eta = 0$, i.e., when the optimality criterion of the exit rule (2.5) is to maximize the expected return on investment.

Data. The study is conducted on the S&P 500 index price data. The daily data on average bid-ask closing prices is from the Center for Research in Security Prices (CRSP) and spans the period from January 1st, 1965 to December 31st, 2016.

3.1 Identifying bubble-like periods in past data

Following the methodology in Zhitlukhin and Ziemba (2016), we consider any market, where the price process exhibits directional changes in its trend, to be a bubble-like market. This notion of the bubble-like market is well in line with thinking of a trader that attempts to benefit from bull and bear market periods by entering into a long, respectively, short position in the underlying.

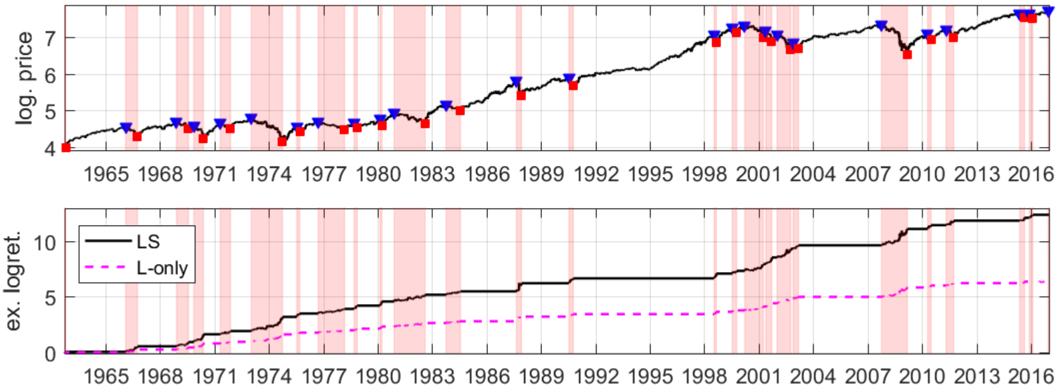
Perception of the market as being in a bull or bear period is *subjective* to such a trader. It is determined by the average holding time of the trader, i.e., for how long on average he keeps a position before liquidating it. The more often this investor would like to trade (i.e., the shorter is the average holding time), the finer is the corresponding price split into bull and bear periods, and the greater is the trader's sensitivity to market fluctuations, or equivalently, the lower is the accepted market volatility.

Appendix B describes the precise procedure to identify bull and bear periods in past price data based on average holding time T_{av} (hereafter, average period duration, or average regime duration). Bull (bear) periods are defined as price realizations between each local price minimum (maximum) and the following local price maximum (minimum). These price locals are subjective to the trader. The main idea behind their identification is that the long-term drift is generally positive in bull periods and negative in bear ones. It changes its sign from positive (negative) to negative (positive) after the price process exhibits a local maximum (minimum) and starts to decrease (increase). Accordingly, a local price maximum (minimum) can be found as the price maximum (minimum) within each period when the drift remains positive (negative). By setting instead two thresholds, one positive and one negative, both different from zero, one can control for frequency of identified locals. The further the thresholds are from zero, the fewer price locals are detected, i.e., the longer is the corresponding average holding time of the trader.

Figure 3.1 illustrates the split of the S&P 500 price data between January 1st, 1965 and December 31st, 2016 into bull and bear periods based on average regime duration T_{av} of 1 year. White and highlighted background areas are, respectively, subjective bull and bear market periods. The bottom panel exhibits cumulative excess (to market) returns of two strategies that adjust their positions according to the prevailing market regime. The LS-strategy holds the S&P 500 long during bull periods and shorts it in bear ones. The L-Only strategy holds the S&P 500 long in bull periods and liquidates it in bear ones with subsequent investment into the risk-free asset with zero interest rate. Both strategies generate positive excess returns.

Table 3.1 reports annualized means and standard deviations of one-period logarithmic returns across all bull and bear periods in the S&P 500, that are identified based on average holding times of 1 month, 3 months, 6 months and 1 year. As expected, we observe that the finer is the data split, the more pronounced are the market drifts of bull and bear periods and the greater is the volatility. Bear periods generally exhibit higher volatility than bull periods.

Figure 3.1: Bull and bear periods in the S&P 500.



NOTES: Bull (white) and bear (pink) periods in the S&P 500 index between January 1st, 1965 and December 31st, 2016 based on average holding time T_{av} of 1 year.
Top panel: The S&P 500 index and its local price maximums (red squares) and minimums (blue triangles), logarithmic scale.
Bottom panel: Cumulative logarithmic returns on LS and L-Only strategies in excess to returns on the S&P 500. The LS (L-Only) strategy holds the S&P 500 long in bull periods and shorts (liquidates) it in bear. Margin requirement of 100%, zero interest rates, no transaction costs.

Positive and negative bubbles. The data split into bull and bear periods allows to identify so-called bubble periods in the price data. We further refer to a combination of a bull and the subsequent bear period as a *positive bubble*, and to a combination of a bear and the following bull period as a *negative bubble*. The data split into positive and negative bubbles is subject to the average holding time T_{av}.

Benchmark exit dates. After the data is split into bull and bear periods, each entry date can be assigned a unique *benchmark exit date*. We define the benchmark exit date as the first moment following the entry date when the underlying changes its regime (from bull to bear and vice versa).

Benchmark exit dates correspond to local price minimums and maximums attributed to a given data split. Intuitively, the benchmark exit date is the optimal moment to close the position that has been formed on the entry date, as it delivers the (locally) highest return.[9]

3.2 Importance of forward looking information (T, p)

Our first study analyses importance of forward looking information (T, p) for timing the bubble burst. Namely, we investigate how varying each of parameters T

[9]Closing the position on the benchmark exit date is more profitable to the trader than on any previous date, as then he can exploit the advantageous mode of the risk factor to the full. Closing after the benchmark exit, on the other hand, would make the trader subject to the disadvantageous risk factor's mode. However, if the position is closed long after the benchmark exit, it might happen that the risk factor is already again in the advantageous mode, and the cumulative realized return is higher than the benchmark return.

Table 3.1: Market regimes for different data splits into bull and bear periods.

$T_{av} \backslash mkt.char.$	μ_p	σ_p	$\lvert\mu_n\rvert$	σ_n
1m.	0.75	13.88	1.13	16.99
3m.	0.44	14.01	0.86	19.39
6m.	0.35	13.87	0.54	19.71
1y.	0.26	13.50	0.45	21.56

NOTES: The reported values are the annualized mean μ_p (μ_n) and standard deviation σ_p (σ_n) of 1-period logarithmic returns across all bull (bear) periods in the S&P 500 between January 1st, 1965 and December 31st, 2016. Bull and bear periods are identified based on average period durations of 1 month, 3 months, 6 months and 1 year.

and p affects performance of the exit rule (2.5) on individual bubbles, given that current and future market regimes (μ_1, σ_1) and (μ_2, σ_2) are known. Intuitively, the two parameters T and p should have the highest impact on the exit rule's signals as they refer to quantitative information on the moment of the bubble burst (peak).

We run the exit rule (2.5) on a set of bubble periods identified ex post in the S&P 500 price data based on average holding times of 1 month, 1 quarter, 6 months and 1 year. Current and future market regimes (μ_1, σ_1) and (μ_2, σ_2) are estimated as mean and standard deviation of 1-period logarithmic returns, respectively, on left and right bubble slopes.[10]

For each bubble period we employ 1 entry date on the left slope of the bubble with

$$t_{entry} = t_{BS} + 0.2 \cdot (t_{BP} - t_{BS}),$$

where t_{BS}, t_{BP} denote, respectively, the start and the peak of the bubble.[11]

For every entry date we consider 5 latest observation dates,

$$t_{latest}(t_{entry}, t_{BP}, x) = t_{entry} + x \cdot (t_{BP} - t_{entry}) \tag{3.1}$$

with multiplier $x = 1, 2, 3, 4$ and 5. The multiplier x determines "closeness" of latest observation date t_{latest} to target benchmark exit date t_{BP} in *relative terms*, i.e., independently of (calendar) duration $t_{BP} - t_{exit}$ of initial market regime (μ_1, σ_1). The cases $x = 1$ and $x \in \{2, 3, 4, 5\}$ differ conceptually. When $x = 1$, the exit rule

[10] Assume that for a given data split a positive (negative) bubble (t_{BS}, t_{BP}, t_{BE}) has been identified. It means that the market is in a bull (bear) period between t_{BS} and t_{BP} (left bubble slope) and in a bear (bull) period between t_{BP} and t_{BE} (right bubble slope); t_{BS}, t_{BE} are local price minimums (maximums) and t_{BP} is a local price maximum (minimum). We note that t_{BP} does not necessarily coincide with the moment of statistical regime change. This is the moment of a directional change in the price trend on a given price realization. Analogously to Zhitlukhin and Ziemba (2016), we employ t_{BP} as an approximation of the moment of statistical regime change.

[11] Considering t_{entry} to the right of the start of a bubble period (i.e., $t_{BS} + 0.2 \cdot (t_{BP} - t_{BS})$ instead of t_{BS}) helps to overcome sample bias of the entry date to always be a local price maximum or minimum.

Table 3.2: Performance of the exit rule (2.5), ex post estimated market regimes.

		stat.	% Bench.					Loss l.max					R. F.Alarm					R. Delay				
T_{av}	N obs.	$mult.\backslash p$	0.2	0.4	0.6	0.8	1	0.2	0.4	0.6	0.8	1	0.2	0.4	0.6	0.8	1	0.2	0.4	0.6	0.8	1
1m.	2110	1	99.92	99.66	98.87	97.56	96.64	0.05	0.21	0.62	1.08	1.14	0.77	3.38	11.70	27.20	43.73	0.00	0.00	0.00	0.00	0.00
		2	96.78	96.99	97.15	97.48	97.80	3.52	3.21	2.99	2.58	2.15	0.22	0.31	0.60	1.95	4.23	58.26	46.54	37.76	28.71	18.77
		3	96.82	96.72	96.86	97.06	97.27	3.96	3.79	3.53	3.21	2.89	0.33	0.33	0.37	0.49	1.05	98.23	78.83	64.46	51.01	38.65
		4	96.78	96.92	96.75	96.84	97.01	4.26	4.03	3.90	3.64	3.32	0.32	0.32	0.32	0.32	0.50	131.84	109.97	89.59	72.27	56.98
		5	96.91	96.78	96.93	96.76	96.86	4.47	4.38	4.14	4.01	3.66	0.32	0.32	0.32	0.32	0.36	168.89	137.30	119.28	95.08	74.99
3m.	1025	1	99.53	99.27	98.38	96.55	94.59	0.33	0.49	0.96	1.71	2.19	2.55	4.53	10.19	21.06	37.09	0.00	0.00	0.00	0.00	0.00
		2	94.59	94.81	94.98	95.30	95.76	5.79	5.47	5.23	4.69	4.09	0.72	1.26	1.43	2.52	4.28	59.54	50.61	42.03	32.68	21.52
		3	94.51	94.51	94.40	94.72	95.01	6.69	6.43	6.24	5.78	5.30	0.46	0.52	0.81	1.26	1.39	103.53	84.12	71.05	58.98	45.90
		4	94.36	94.78	94.46	94.63	94.64	7.50	7.01	6.90	6.45	6.07	0.22	0.47	0.52	0.80	1.23	137.84	119.45	97.21	83.88	66.77
		5	94.27	94.28	94.75	94.54	94.61	8.11	7.82	7.27	7.03	6.56	0.21	0.29	0.47	0.52	0.81	174.11	145.74	128.35	106.16	86.97
6m.	515	1	98.47	97.84	96.93	95.13	92.83	0.73	0.90	1.37	2.20	2.92	4.80	7.69	12.99	20.90	33.12	0.00	0.00	0.00	0.00	0.00
		2	93.21	93.29	93.47	93.17	93.76	7.09	6.78	6.52	6.46	5.83	3.05	3.72	4.18	5.53	6.50	56.85	49.06	43.48	34.22	24.92
		3	93.76	93.96	94.07	93.93	93.69	8.79	8.35	7.98	7.63	7.18	2.13	2.89	3.13	3.76	4.14	94.39	81.38	73.94	61.54	52.70
		4	93.11	93.31	93.57	94.16	94.42	10.15	9.42	9.07	8.51	8.10	1.14	2.71	2.84	3.10	3.71	123.70	105.96	95.18	87.66	73.61
		5	92.70	93.25	93.18	93.15	93.57	10.70	10.16	9.66	9.58	8.93	1.13	1.15	2.75	2.87	3.12	153.99	130.59	113.47	103.80	91.68
1y.	245	1	96.77	95.89	93.68	92.97	90.02	1.04	1.48	2.68	2.93	3.64	6.33	8.72	15.13	18.87	30.96	0.00	0.00	0.00	0.00	0.00
		2	89.15	89.57	89.62	90.00	90.29	9.06	8.65	8.51	8.06	7.67	5.87	6.21	6.21	6.29	7.49	48.42	43.03	37.84	32.70	22.60
		3	89.29	89.37	88.54	88.98	89.62	11.10	10.68	10.08	9.67	9.04	2.67	4.50	5.91	6.20	6.21	84.63	74.48	60.06	54.29	48.02
		4	88.34	88.70	88.87	88.29	88.88	13.05	12.02	11.69	10.81	10.01	1.64	2.43	3.15	5.78	5.85	104.78	99.20	86.18	72.47	61.72
		5	87.62	87.89	88.23	88.22	87.71	13.96	13.45	12.46	12.28	11.33	1.62	1.64	2.43	3.12	5.76	129.92	110.87	106.04	95.45	80.78

NOTES: The reported numbers are the average ratio of return on the exit rule to the maximum (benchmark) return in percent, the average lost return of the exit rule relative to the past local maximum in percent and the average false alarm and delay relative to the duration of regime (μ_1, σ_1) in percent.
Positive and negative bubbles are identified for daily price data on the S&P 500 index between January 1st, 1965 and December 31st, 2016, based on average period durations of 1 month, 3 months, 6 months and 1 year. Market regimes (μ_1, σ_1) and (μ_2, σ_2) are estimated ex post from left and right bubble slopes as mean and standard deviation of 1-period logarithmic returns.

Table 3.3: Performance of the exit rule (2.5), mistakes in market regimes.

		stat.	% Bench.					Loss l.max					R. F.Alarm					R. Delay				
T_{av}	N obs.	*mult.*\p	0.2	0.4	0.6	0.8	1	0.2	0.4	0.6	0.8	1	0.2	0.4	0.6	0.8	1	0.2	0.4	0.6	0.8	1
1m.	2110	1	99.42	98.94	98.31	97.35	96.48	0.23	0.45	0.71	0.97	1.05	6.66	11.82	20.04	33.50	48.07	0.00	0.00	0.00	0.00	0.00
		2	97.02	97.11	97.25	97.46	97.70	3.28	3.07	2.80	2.47	2.00	1.84	2.96	3.95	6.24	10.29	58.09	48.32	39.34	29.23	18.17
		3	96.63	96.73	96.84	97.01	97.22	3.83	3.63	3.39	3.12	2.76	1.80	2.15	3.13	3.85	5.22	95.96	81.57	68.42	57.27	41.78
		4	96.83	96.91	96.97	97.02	97.20	4.08	3.89	3.72	3.50	3.18	1.13	1.63	2.40	2.91	3.49	139.47	118.34	102.30	84.28	65.48
		5	96.96	97.01	97.12	97.10	97.11	4.43	4.24	3.98	3.81	3.52	1.02	1.14	1.55	1.82	2.42	179.84	155.82	134.00	113.01	88.20
3m.	1025	1	98.31	97.71	97.01	96.01	94.12	0.60	0.85	1.17	1.58	1.94	11.13	15.80	21.84	30.74	44.30	0.00	0.00	0.00	0.00	0.00
		2	94.83	94.88	94.82	95.08	95.40	5.02	4.84	4.55	4.12	3.62	6.89	8.02	9.77	10.77	14.11	56.26	47.25	38.55	28.40	19.68
		3	95.12	94.89	94.79	94.99	95.16	5.86	5.72	5.45	5.14	4.61	5.08	6.25	7.24	8.15	9.44	97.38	84.54	75.42	63.86	46.99
		4	95.15	95.28	95.33	95.34	95.26	6.36	6.17	5.94	5.69	5.41	6.41	6.82	8.07	8.86	9.49	133.30	120.24	110.34	97.66	78.81
		5	95.50	95.53	95.70	95.47	95.07	7.09	6.91	6.62	6.27	6.24	6.04	6.46	6.94	7.32	7.95	168.21	146.86	137.21	122.20	98.51
6m.	515	1	96.07	95.18	94.16	93.45	92.10	1.23	1.58	1.96	2.15	2.21	17.31	21.85	27.44	31.97	44.05	0.00	0.00	0.00	0.00	0.00
		2	92.10	92.25	92.49	93.03	92.93	6.38	6.08	5.61	4.94	4.54	12.13	13.72	15.60	17.09	20.65	45.02	38.23	33.51	27.43	19.32
		3	90.73	90.99	90.81	91.22	91.47	7.66	7.35	7.17	6.66	6.15	11.02	11.61	13.91	14.89	17.11	79.55	74.09	65.81	56.06	44.74
		4	90.33	90.73	90.79	91.19	91.25	8.85	8.13	7.94	7.26	7.04	10.07	10.98	11.47	13.72	14.10	111.18	95.88	92.42	79.72	67.46
		5	91.74	91.34	91.51	90.97	91.12	9.64	9.27	8.81	8.69	8.02	6.01	7.67	8.04	9.31	10.50	134.02	117.59	101.91	86.87	78.00
1y.	245	1	92.04	91.07	89.76	88.83	87.28	1.59	1.94	2.69	3.04	3.32	25.00	28.39	32.40	38.59	47.10	0.00	0.00	0.00	0.00	0.00
		2	90.31	90.58	90.53	90.48	89.53	8.25	7.66	7.51	7.25	6.91	9.46	11.81	13.20	13.95	17.23	42.43	35.83	32.24	28.50	21.18
		3	89.67	90.18	90.56	90.63	90.19	10.19	9.39	8.72	8.53	8.29	7.39	9.33	11.50	11.81	13.24	77.10	68.60	61.02	51.30	42.61
		4	91.81	92.05	92.14	92.45	92.98	10.69	10.27	10.06	9.84	9.01	6.59	8.18	9.24	9.56	10.10	95.04	87.68	83.00	74.59	60.31
		5	92.24	92.66	92.84	92.81	93.38	12.20	11.24	10.93	10.69	10.22	6.41	6.57	8.17	9.23	9.39	111.00	105.01	99.95	95.79	80.99

NOTES: The reported numbers are the average ratio of return on the exit rule to the maximum (benchmark) return in percent, the average lost return of the exit rule relative to the past local maximum in percent and the average false alarm and delay relative to the duration of regime (μ_1, σ_1) in percent.
Positive and negative bubbles are identified for daily price data on the S&P 500 index between January 1st, 1965 and December 31st, 2016, based on average period durations of 1 month, 3 months, 6 months and 1 year. Market regimes (μ_1, σ_1) and (μ_2, σ_2) are estimated ex post from left and right bubble slopes as mean and standard deviation of 1-period logarithmic returns and then perturbed by random mistakes of up to 50% percent.

is expected to signal on the latest observation date, as no change of regime occurs within the observation period. On the other hand, when $x \geq 2$, the exit rule is expected to signal soon after t_{BP}.

Probability p takes values in $\{20\%, 40\%, 60\%, 80\%, 100\%\}$.

The results are given in Table 3.2. The reported numbers are the average ratio of return on the exit rule to the benchmark return in percent, the average lost return of the exit rule relative to the past local maximum in percent, and the average relative false alarm and delay in percent.[12] The data is sorted based on the multiplier x (equation (3.1)) and the probability p. The averages are computed within each multiplier-probability class. Table 3.3 outputs the same statistics for the case of ex post estimated market regimes being perturbed by random and independent mistakes of up to 50%.

We find that the exit rule is sensitive to both the observation window T (multiplier x) and the probability p. The average relative delay varies from 0% to more than 150%, the average relative false alarm is between 0.2% and 45% and the average lost return changes from less than 0.5% to more than 13%, depending on the data split. We observe three pronounced patterns of the exit rule's performance, all robust to mistakes in market regimes.

Firstly, when the benchmark exit date coincides with the end of the observation period, i.e., $x = 1$, performance of the exit rule, as measured by the average percentage of the benchmark return, the average lost return relative to the past local maximum as well as the average relative false alarm, decreases with the probability p of regime change. The result is intuitive: The benchmark exit date is the global maximum within the observation period, hence the later is the signal, the better is the performance of the exit rule. At the same time, the lower is the probability p, the less sensitive is the exit rule to price fluctuations, hence the later is the signal.

Secondly, when the benchmark exit is within the observation period (i.e., the multiplier $x > 1$), the exit rule's performance, as measured by the average lost return relative to the past local maximum as well as the average relative delay and the average relative false alarm, increases with the certainty of beliefs p. For all configurations of (T_{av}, x) we observe that the average delay is significantly longer than the average false alarm, implying that the exit rule consistently "overshoots" the benchmark exit date. Increasing the probability p of regime change within the same observation window makes the exit rule react to market corrections faster.

Thirdly, for each given level of p, we observe that the closer is the latest obser-

[12] The *return on the signal from the exit rule* is defined as $ret_{exit} = S_{exit}/S_{entry}$ with S_{exit} (S_{entry}) the price of the underlying on the exit (entry) date. The *benchmark return* is $ret_{BP} = S_{BP}/S_{entry}$, where S_{BP} is the local price maximum (minimum) of a positive (negative) bubble. The *lost return relative to the past local maximum* is $ret_{lost} = (S_{l.max} - S_{exit})/S_{l.max} = 1 - ret_{exit}/ret_{l.max}$ with $S_{l.max} = \max\{S_t \mid t \in [t_{entry}, t_{exit}]\}$ for positive bubbles and $S_{l.max} = \min\{S_t \mid t \in [t_{entry}, t_{exit}]\}$ for negative bubbles. The returns are computed in assumption of zero transaction costs. The *relative false alarm (delay)* is the ratio of the false alarm (delay) in calendar days, $\max(t_{BP} - t_{exit}, 0)$ ($\max(t_{exit} - t_{BP}, 0)$), to $t_{BP} - t_{entry}$ in calendar days.

Table 3.4: Distribution of optimal (T, p) when benchmark exit is outside of observation window.

opt. crit.	% Bench.					Loss l.max				
$T_{av} \backslash p$	0.2	0.4	0.6	0.8	1	0.2	0.4	0.6	0.8	1
1m.	34.71	32.27	23.36	9.08	0.59	32.83	30.52	22.26	9.70	4.69
	(35.00)	(29.08)	(21.24)	(10.94)	(3.74)	(31.80)	(26.86)	(20.25)	(12.38)	(8.70)
3m.	31.38	29.76	24.39	13.33	1.14	30.52	28.93	23.69	13.20	3.66
	(29.88)	(26.21)	(22.87)	(15.86)	(5.18)	(28.21)	(25.64)	(22.28)	(16.03)	(7.85)
6m.	31.63	29.59	22.79	13.61	2.38	30.10	28.76	22.07	14.05	5.02
	(32.72)	(25.37)	(20.59)	(14.71)	(6.62)	(28.13)	(24.31)	(20.14)	(15.28)	(12.15)
1y.	31.88	29.71	18.84	16.67	2.90	30.07	27.97	18.18	17.48	6.29
	(29.50)	(24.46)	(19.42)	(15.83)	(10.79)	(28.06)	(23.74)	(20.14)	(15.83)	(12.23)

NOTES: Distribution of optimal (T, p) for data splits based on average regime durations T_{av} of 1 month, 3 months, 6 months and 1 year and for two optimality criteria, the ratio of return on the exit rule to the benchmark return (% Bench.) and the lost return of the exit rule relative to the past local maximum (Loss. l.max). The reported values are frequencies in percent of each pair (T, p) to be optimal for given T_{av} and optimality criterion. Values in round brackets are the corresponding frequencies for the case of mistakes in market regimes.

vation date to the benchmark exit date, the better is the exit rule's performance, as measured by the average lost return relative to the past local maximum. At the same time, the lower is the average relative delay and the higher is the average false alarm, implying that the rule signals earlier on average.

Distribution of optimal (T, p). Tables 3.4 and 3.5 report distributions of optimal (T, p) for the case when the benchmark exit is outside, respectively, within the observation window. The two optimality criteria are the ratio of return on the exit rule to the benchmark return and the lost return of the exit rule relative to the past local maximum.

The results generally confirm conclusions of the previous section. We also find that influence of probability p diminishes with the observation horizon. While the distribution is monotonous on p for $x = 2$, it flattens for $x = 5$ (Table 3.5). Analogously, influence of T increases with certainty of beliefs p.

Value of optimal (T, p). Table 3.6 quantifies the value of optimal (T, p) for performance of the exit rule. Thus, the optimal values increase the ratio of return on the exit rule to the benchmark return in the range from 4% to 11% percent on average, depending on the data split. They also reduce the lost return relative to the past local maximum by 3.5% – 9.5% on average.

Table 3.5: Distribution of optimal (T, p) when benchmark exit is within observation window.

	opt. crit.	% Bench.					Loss l.max				
T_{av}	$mult.\backslash p$	0.2	0.4	0.6	0.8	1	0.2	0.4	0.6	0.8	1
1m.	2	4.07	5.15	6.54	10.10	17.57	4.63	5.59	6.81	10.12	19.12
		(4.61)	(4.77)	(5.73)	(7.77)	(12.27)	(4.83)	(4.94)	(6.00)	(8.44)	(15.02)
	3	3.25	3.35	3.71	4.64	6.49	3.41	3.46	3.81	4.88	6.56
		(3.86)	(3.80)	(3.70)	(4.98)	(6.32)	(3.72)	(3.50)	(3.72)	(4.72)	(6.53)
	4	3.25	3.04	2.89	3.09	4.02	3.00	2.95	2.75	3.20	4.12
		(3.91)	(3.59)	(3.75)	(3.86)	(4.34)	(3.56)	(3.50)	(3.87)	(4.25)	(4.56)
	5	4.89	3.92	3.66	3.19	3.19	3.71	2.95	3.00	2.90	3.05
		(5.14)	(4.82)	(4.50)	(4.34)	(3.91)	(3.77)	(3.72)	(3.72)	(3.66)	(3.98)
3m.	2	4.43	5.14	5.67	10.46	19.33	5.68	5.86	6.22	11.72	21.85
		(4.29)	(5.22)	(5.97)	(7.84)	(12.13)	(5.04)	(5.37)	(6.83)	(7.32)	(12.03)
	3	3.19	3.19	3.19	4.08	5.85	3.55	3.73	3.55	3.91	5.68
		(4.85)	(3.36)	(2.61)	(3.92)	(6.34)	(3.90)	(3.90)	(3.25)	(4.39)	(6.99)
	4	4.43	3.72	2.84	2.84	3.72	3.55	3.20	2.49	2.66	3.73
		(3.92)	(3.54)	(3.36)	(2.80)	(4.10)	(3.90)	(3.58)	(3.41)	(4.23)	(4.55)
	5	4.08	4.43	4.26	2.66	2.48	2.13	2.66	3.20	2.31	2.31
		(5.60)	(5.97)	(5.41)	(4.48)	(4.29)	(3.74)	(4.07)	(4.55)	(4.39)	(4.55)
6m.	2	5.71	5.71	5.11	7.21	15.32	6.71	6.71	6.12	7.87	15.74
		(5.97)	(5.60)	(7.09)	(6.34)	(10.07)	(4.86)	(5.56)	(6.94)	(6.94)	(9.72)
	3	4.20	3.90	3.60	4.20	4.20	4.08	3.79	3.50	4.08	4.37
		(2.99)	(3.73)	(3.36)	(4.10)	(5.22)	(4.51)	(4.17)	(3.82)	(4.17)	(6.25)
	4	4.80	4.80	3.60	3.30	3.60	3.79	4.37	4.08	3.79	3.50
		(2.99)	(3.73)	(2.99)	(2.99)	(4.10)	(2.43)	(3.47)	(2.78)	(3.13)	(4.17)
	5	4.20	4.50	4.80	3.60	3.60	3.21	3.21	3.79	3.50	3.79
		(6.34)	(4.48)	(5.97)	(5.22)	(6.72)	(3.13)	(4.51)	(5.21)	(6.94)	(7.29)
1y.	2	5.41	4.50	3.60	11.71	21.62	6.48	5.56	4.63	14.81	25.93
		(3.25)	(4.88)	(8.13)	(8.94)	(11.38)	(4.86)	(6.94)	(5.56)	(8.33)	(11.81)
	3	4.50	4.50	0.90	1.80	2.70	2.78	3.70	2.78	2.78	3.70
		(4.07)	(1.63)	(1.63)	(0.81)	(5.69)	(4.17)	(2.78)	(4.86)	(4.17)	(3.47)
	4	4.50	5.41	4.50	2.70	0.90	2.78	3.70	2.78	2.78	2.78
		(4.88)	(4.88)	(5.69)	(5.69)	(7.32)	(2.78)	(4.17)	(5.56)	(6.25)	(6.94)
	5	4.50	4.50	4.50	4.50	2.70	1.85	1.85	2.78	3.70	1.85
		(5.69)	(3.25)	(3.25)	(4.07)	(4.88)	(3.47)	(2.08)	(2.78)	(4.17)	(4.86)

NOTES: Distribution of optimal (T, p) for data splits based on average regime durations T_{av} of 1 month, 3 months, 6 months and 1 year and for two optimality criteria, the ratio of return on the exit rule to the benchmark return (% Bench.) and the lost return of the exit rule relative to the past local maximum (Loss. l.max). The reported values are frequencies in percent of each pair (T, p) to be optimal for given T_{av} and optimality criterion. Values in round brackets are the corresponding frequencies for the case of mistakes in market regimes.

Table 3.6: Value of optimal (T, p).

opt. crit.	% Bench.				Loss l.max			
$T_{av}\backslash\Delta stat.$	% Bench.	Loss l.max	R. F.Alarm	R. Delay	% Bench.	Loss l.max	R. F.Alarm	R. Delay
1m.	4.51	-3.31	0.53	-0.28	3.74	-3.65	1.72	-30.51
	(4.75)	(-2.99)	(-3.28)	(14.96)	(3.40)	(-3.40)	(1.91)	(-31.28)
3m.	6.59	-5.18	-1.76	-18.24	5.47	-5.56	1.05	-45.48
	(7.80)	(-4.24)	(-5.38)	(1.03)	(4.75)	(-4.43)	(3.97)	(-35.13)
6m.	7.28	-5.92	-0.22	-30.47	4.12	-6.84	0.87	-58.79
	(9.12)	(-5.50)	(-5.55)	(-34.15)	(4.70)	(-6.41)	(2.93)	(-61.36)
1y.	11.53	-8.38	-1.98	-35.29	8.52	-9.48	0.16	-54.48
	(9.65)	(-5.90)	(-10.22)	(-16.63)	(2.84)	(-5.67)	(9.19)	(-36.47)

NOTES: The reported numbers are the average differences between statistics computed for the exit rule with the optimal (T, p) and their reference values computed with $T = 3 \cdot (t_{BP} - t_{entry})$ and $p = 80\%$. The optimality criteria are the ratio of return on the exit rule to the benchmark return and the lost return of the exit rule relative to the past local maximum. Values in round brackets correspond to the case of mistakes in market regimes.

3.3 Importance of information on market regimes

Tables 3.4, 3.5 and 3.6 suggest that mistakes in market regimes have no effect on the dependence of performance of the exit rule on forward looking parameters T and p.

Value of information on market regimes. Table 3.7 quantifies the value of information on market regimes for performance of the exit rule. The reported numbers are the average differences between four statistics for the two cases when the exit rule is specified by ex post estimated market regimes with, respectively, without mistakes. Values are sorted based on the size of relative mistakes in each of four characteristics μ_1, σ_1, μ_2 and σ_2.

Unlike the case of optimal (T, p), we find no directional effect of mistakes on either the ratio of return on the exit rule to the benchmark return or the lost return of the exit rule relative to the past local maximum. All types of mistakes increase the average false alarm, but exhibit no clear pattern for the average delay. The latter implies an increase in volatility of signals of the exit rule after mistakes were added.

Comparing the results to that in Table 3.6, we find that mistakes in market regimes have significantly less pronounced effect on the exit rule's performance than the observation window T and the probability of regime change p.

How severe a market correction should be for the exit rule to signal? The lost return relative to the past local maximum characterizes how big a price

Table 3.7: Value of information on market regimes.

	$\Delta stat.$	% Bench.				Loss l.max				R. F.Alarm				R. Delay			
T_{av}	$\Delta, \% \setminus mkt.char.$	μ_1	μ_2	σ_1	σ_2	μ_1	μ_2	σ_1	σ_2	μ_1	μ_2	σ_1	σ_2	μ_1	μ_2	σ_1	σ_2
1m.	$(25, 50)$	-0.22	-0.08	-0.09	0.05	-0.19	-0.27	0.41	-0.31	6.91	4.66	-0.10	3.88	-3.63	0.68	28.88	-2.10
	$(0, 25)$	-0.08	-0.02	0.07	-0.12	-0.16	-0.12	0.17	-0.19	3.23	3.82	0.27	2.83	6.68	3.14	22.55	-3.65
	$(-25, 0)$	0.21	0.04	0.06	-0.12	-0.07	-0.15	-0.08	-0.11	1.05	2.75	0.29	4.33	14.74	6.17	-0.11	1.27
	$(-50, -25)$	-0.10	-0.13	-0.23	0.00	0.06	0.15	-0.98	0.20	2.34	2.37	14.29	2.53	6.75	14.72	-29.44	27.51
3m.	$(25, 50)$	1.07	0.71	1.47	0.64	-0.84	-0.73	-0.05	-0.75	8.55	12.20	1.74	8.14	-6.88	3.48	18.95	-7.82
	$(0, 25)$	0.89	-0.66	-0.71	0.24	-0.72	-0.46	0.25	-0.68	4.80	7.04	0.60	3.91	9.92	-5.55	16.44	-7.11
	$(-25, 0)$	-0.44	0.46	-0.18	0.58	-0.70	-0.62	-0.39	-0.56	7.28	5.50	2.02	5.92	-4.38	11.05	1.90	10.29
	$(-50, -25)$	-0.41	0.67	0.69	-0.66	-0.34	-0.78	-2.92	-0.52	6.88	3.56	28.52	9.13	18.30	9.60	-24.94	30.02
6m.	$(25, 50)$	-2.43	-8.33	-8.21	-0.08	0.65	-1.26	0.49	-1.00	14.64	15.26	9.71	8.99	-9.18	-25.66	-25.25	3.38
	$(0, 25)$	-0.57	0.45	-0.11	-2.16	0.27	-0.06	-0.55	0.06	5.18	7.56	2.57	10.73	4.64	3.69	28.17	-15.55
	$(-25, 0)$	-0.15	-0.47	0.45	-8.81	-1.78	-1.81	-1.04	-1.32	6.56	11.45	3.45	18.54	1.45	5.47	4.88	-31.86
	$(-50, -25)$	-6.16	-1.99	-1.54	1.25	-2.44	-0.59	-3.17	-1.88	16.06	9.64	29.23	4.72	-14.88	-4.01	-26.73	31.79
1y.	$(25, 50)$	1.24	2.15	1.39	2.75	1.80	-2.36	-1.26	-0.29	-4.20	3.36	5.93	5.11	17.77	-17.30	6.35	19.29
	$(0, 25)$	-0.17	-0.25	1.63	0.59	-2.62	-1.94	1.23	-3.80	14.25	15.25	0.94	14.70	-17.47	-16.94	18.40	-32.67
	$(-25, 0)$	4.61	1.56	2.30	3.03	-3.50	-1.90	-1.64	-1.41	3.87	6.06	-4.56	-4.28	-8.46	-5.24	-4.39	-15.39
	$(-50, -25)$	1.43	2.88	1.29	-0.17	-0.94	1.70	-2.67	0.72	7.67	-1.01	18.10	3.52	-4.89	28.56	-26.79	2.44

NOTES: The reported numbers are the average differences between statistics computed for the exit rule with mistakes in market regimes and their reference values for ex post estimated market regimes. The considered statistics are the ratio of return on the exit rule to the benchmark return in percent, the lost return of the exit rule relative to the past local maximum in percent, the relative false alarm and delay in percent.

correction should be for the exit rule to signal. Thus, Tables 3.2 and 3.3 suggest a necessary average price correction between 3% and 10%, depending on data split. We find no significant difference in sizes of required price drops and recoveries when considering positive and negative bubbles separately.

3.4 Summary of the main findings

Our sensitivity analysis suggests that the forward looking observation window T and the probability p of regime change determine most of variability of the exit rule's signals. The relation between the optimal T given p and vice versa depends on whether the benchmark exit date is within the observation window T. For practical applications this implies that one should focus on joint calibration of parameters (T, p), whereas market regimes can be estimated, e.g., from past data. The optimal specification of the exit rule reduces the price correction required for the exit rule to signal by 3%–10% on average.

4 Dynamic application of the exit rule

In this section we examine out-of-sample performance of the investment strategy that makes use of the exit rule's signals to enter and exit the market.

Data. The study is conducted on the same price data on the S&P 500 index as the sensitivity analysis in Section 3. To model the risk-free asset we employ daily data on annualized yields on 3-months treasury Bills from the FRED Economic Data (Federal Reserve Bank of St. Louis Economic Data). We apply 50 basis points transaction costs and use the margin requirement of 100% for a short position. Asset loan is free of charge, but deposited initial margins and short sale proceeds earn no interest.

Repetitive structure of the market. We extend the stochastic disorder model (2.1) to multiple regime changes with 2 market regimes of 1-period logarithmic returns:

$$X_t = \begin{cases} \mu_p + \sigma_p \cdot \xi_t, & \delta_t = 1, \\ \mu_n + \sigma_n \cdot \xi_t, & \delta_t = -1. \end{cases}$$

The unobservable process δ_t determines whether the market is in a bull, (μ_p, σ_p), or bear, (μ_n, σ_n), mode.

As before, we assume that perception of market regimes is subjective to a trader. For a trader with average holding time T_{av} we approximate values μ_p, μ_n and σ_p, σ_n by means and standard deviations of 1-period logarithmic returns across all subjective bull, respectively, bear price periods (refer to Appendix B for details).

Investment strategy. The strategy makes use of the repetitive structure of the market. It employs signals from the exit rule to identify the prevailing market regime (bull or bear). We consider 2 types of the entry-exit strategy. The "LS"-strategy holds the S&P 500 long during identified bull periods and shorts it during bear ones. The "L-Only" strategy invests everything into the risk-free asset during bear periods.

The entry-exit strategy requires calibration to past data to identify the market regime on the first trading day as well as to specify the two exit rules that it employs.[13] Market characteristics (μ_p, σ_p) and (μ_n, σ_n) are estimated subject to average period duration T_{av} (characteristic of a trader). The market regime (bull or bear) of the first trading date is approximated by the market regime on the last date of the calibration period. Parameters (T_p, p_p) and (T_n, p_n) are selected as to maximize the total return on the strategy during the calibration period.

Once calibrated, the strategy runs the exit rule to identify the moment of upcoming change of regime. If the exit rule signals before the latest observation date, the strategy updates the information on the current regime (from bull to bear and vice versa) and sets up a new exit rule, which corresponds to the just identified regime. If the exit rule signals on the latest observation date, the signal is ignored and the same exit rule is run again. The procedure is repeated until the end of the testing period.

4.1 Performance of the entry-exit investment strategy

Table 4.1 reports performance of the entry-exit investment strategy between January 1st, 1975 and December 31st, 2016. The reported values are the total annualized return in percent and the Sharpe ratio of the strategy. The buy-and-hold strategy (B-H) is used as the benchmark strategy.

We consider four specifications of the entry-exit strategy, each corresponding to a particular value of parameter T_{av} employed for data split in estimation of two market regimes. The market regimes as well as parameters (T_p, p_p), (T_n, p_n) are first calibrated using the data between January 1st, 1965 and December 31st, 1974, and then recalibrated every $10 * T_{av}$ observation days using the whole sample of past data up to the current date. Probabilities p_p, p_n are optimized over the set of values $\{20\%, 40\%, 60\%, 80\%, 100\%\}$. Observation window T_p (T_n) is optimized over the set $x \cdot \widetilde{T}$, where the multiplier x takes values in $[0.5, 5]$ with linear grid size of 0.5, and $\widetilde{T}$ denotes the average duration of bull (bear) periods identified while estimating market regimes.

We find that all 8 entry-exit strategies generate positive returns in-sample, but these returns decrease or even become negative when the strategy is run out-of-

[13] The exit rule parametrized by (μ_p, σ_p), (μ_n, σ_n), (T_p, p_p) is used to identify a change of market regime from bull to bear. The exit rule parametrized by (μ_n, σ_n), (μ_p, σ_p), (T_n, p_n) is used to identify a change of market regime from bear to bull.

Table 4.1: Out-of-sample performance of the entry-exit investment strategy in 1975-2016.

period			Nov 1980 – Aug 1982		Aug 1987 – Dec 1987		Mar 2000 – Oct 2002		Oct 2007 – Mar 2009		Jan 1975 –Dec 2016	
STR	T_{av}	*type\stat.*	Ret.	Sh.R.	Ret.	Sh.R.	Ret.	Sh.R.	Ret.	Sh.R.	Ret.	Sh.R.
entry-exit	1m.	LS	-27.78	-2.719	23.58	0.563	-33.68	-1.809	-30.76	-1.000	-7.04	-0.613
											(1.33)	(-0.100)
		L-Only	-14.18	-2.063	-40.87	-3.076	-23.35	-1.696	-33.40	-2.043	3.36	-0.025
											(4.51)	(0.057)
	3m.	LS	-17.20	-1.979	-11.82	-0.271	-32.75	-1.694	-12.17	-0.253	1.51	-0.096
											(4.53)	(0.075)
		L-Only	-13.15	-1.911	-18.69	-1.460	-17.47	-1.254	-35.96	-2.314	5.24	0.109
											(5.50)	(0.309)
	6m.	LS	-18.94	-2.141	45.68	0.980	-32.94	-1.746	-15.02	-0.331	-1.53	-0.272
											(3.92)	(0.038)
		L-Only	-7.99	-1.636	-4.54	-0.558	-23.00	-1.849	-23.86	-1.837	4.78	0.072
											(7.41)	(0.273)
	1y.	LS	-6.57	-1.203	21.89	0.533	-29.35	-1.569	-3.60	-0.033	0.20	-0.171
											(3.22)	(-0.004)
		L-Only	-4.68	-1.314	-8.35	-1.039	-20.31	-1.926	-5.33	-0.859	5.19	0.102
											(7.66)	(0.316)
B-H			-3.47	-0.999	-45.93	-1.193	-15.29	-0.742	-35.71	-1.001	8.54	0.295

NOTES: The reported numbers are the annualized total returns in percent and the Sharpe ratios of the entry-exit and buy-and-hold (B-H) investment strategies. T_{av} is the average regime duration used for data split to estimate market regimes employed by the entry-exit strategy. Values in round brackets correspond to statistics computed in-sample.

sample. Low Sharpe ratios indicate high volatilities of generated returns. We also observe that neither of the entry-exit strategies can outperform the buy-and-hold strategy over the whole period 1965–2016 either in- or out-of-sample.

Table 4.1 also reports performance of the strategies in four historical bear markets during the testing period. We observe that the entry-exit strategy outperforms the buy-and-hold strategy both during the Crash of 1987 and after the dot-com bubble in 2000–2002, for all specifications and types. Even though in some cases the returns on the entry-exit strategy are negative, the strategy does generate excess to market returns.

Figure 4.1 plots signals from the "LS"-entry-exit strategy with T_{av} of 1 year between January 1980 and December 2000. Highlighted areas denote bear market periods as identified by the strategy. We observe that the entry-exit strategy can generate positive returns in excess to the buy-and-hold strategy, if it identifies the prevailing market trend correctly. In particular, the strategy timely detects the Crash of 1987 and enters into a corresponding short position before the market collapses. On the other hand, the strategy fails to detect the following directional change in the market trend, and looses half of its realized returns during the after crash market recovery.

Further, starting 1997 the strategy misidentifies the prevailing market regime and enters into short positions during bull periods and long positions during bear ones. This misalignment could not be corrected until the spring of 2003 and resulted in persistent losses of the strategy. We recall that the strategy was recalibrated in January 1995 and then only in January 2005, therefore underperformance of the strategy during 1997–2003 might be due to inappropriate exit rule's specifications. In particular, because of a long bull period between 1991 and 1997, it could be that the strategy was set to have low tolerance to market volatility. Hence, it reacted many times to market movements in 1997–2003, consistently loosing not only because of market regime's misspecification but also on transaction costs.[14]

Figure 4.2 plots signals from the "LS"-entry-exit strategy with T_{av} of 3 months between January 1980 and December 2009 as well as between January 1985 and December 1990. This strategy also signals before the market collapse in 1987. It manages to detect the following market recovery sooner than the strategy based on T_{av} of 1 year, but nevertheless looses potential returns by producing misleading signals in late autumn 1987 – winter 1988. Further, in 2008–2009 the strategy detects correctly the negative market trend, which helped it to outperform the buy and hold strategy in this period.

The example of the Crash of 1987 suggests that the exit rule might serve as an insurance against significant market corrections. Figures 4.1 and 4.2 show that the exit rule reacts in a timely manner to sharp price declines, characteristic of market

[14]Because of technical challenges this issue is left for future research.

crashes but not recoveries.

Long delays in the exit rule's signals to identify directional changes from negative to positive market trends complicate its application as a single investment instrument.

5 Conclusion and outlook

This paper tests in- and out-of-sample performance of the exit rule by Shiryaev and Zhitlukhin (2012a), Shiryaev et al. (2014, 2015) and Zhitlukhin and Ziemba (2016). We find that signals of the exit rule are very fragile, particularly in response to varying the forward looking observation window. Our dynamic setting tests of the exit rule suggest that it cannot be employed as a single investment instrument, but has potential as an insurance against market corrections and crashes.

There are several aspects that go beyond the scope of this paper. Firstly, because of technical challenges we could not recalibrate the entry-exit strategy sequentially in response to new market information. In particularly, the entry-exit strategy with T_{av} of 1 year was recalibrated only once in every 10 years. To this regard, it would be interesting to investigate whether performance of the strategy would improve, and if so, to which extent, if it were possible to recalibrate it more regularly.

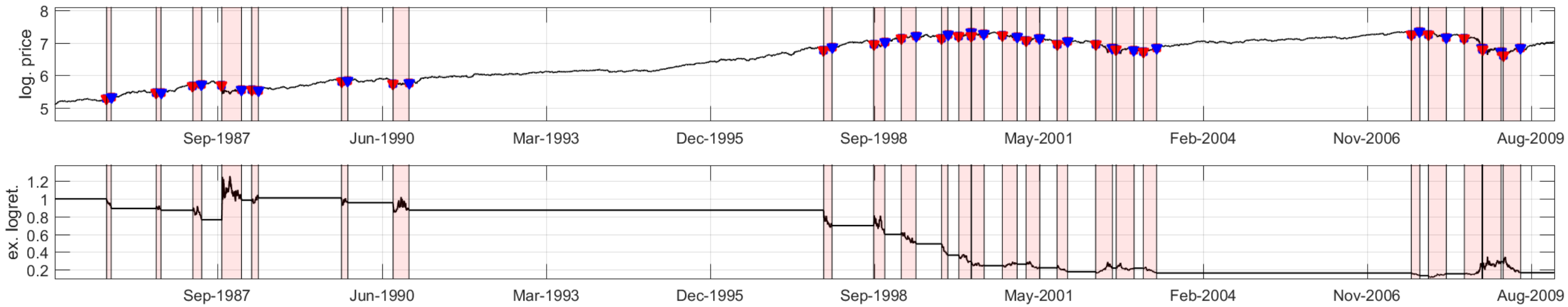
Secondly, the entry-exit strategy does not take into account the posterior probability of the market to be in a bull or bear period conditional on all past prices, as the exit rule employs only the price data starting from the beginning of its observation period. It might be that coordinating the signal from the exit rule with the posterior probability of the market regime conditional on all past price data (as e.g., in Honda (2003)) would fasten alignment of the strategy's position with the actual market regime in case of misalignment.

Thirdly, one could explore applications of the exit rule in conjuncture with other market predictors, that provide exogenous (to the exit rule) information on a reasonable observation window and the corresponding probability of correction. To this aim, one could employ, e.g., real-time diagnostics of bubble presence by Ardila, Sanadgol, Cauwels, and Sornette (2017), Zhang et al. (2016), Watanabe, Takayasu, and Takayasu (2007) or predictive statistics from Koivu, Pennanen, and Ziemba (2005), Lleo and Ziemba (2016).

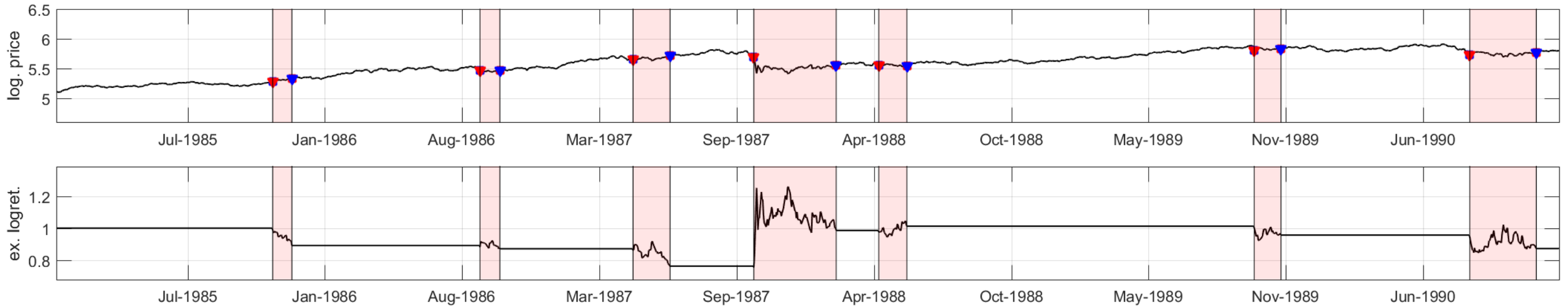
References

Ardila, D., D. Sanadgol, P. Cauwels, and D. Sornette (2017). Identification and critical time forecasting of real estate bubbles in the USA. *Quantitative Finance 17*(4), 613–631.

Figure 4.1: Out-of-sample performance of the entry-exit investment strategy with T_{av} of 1 year.



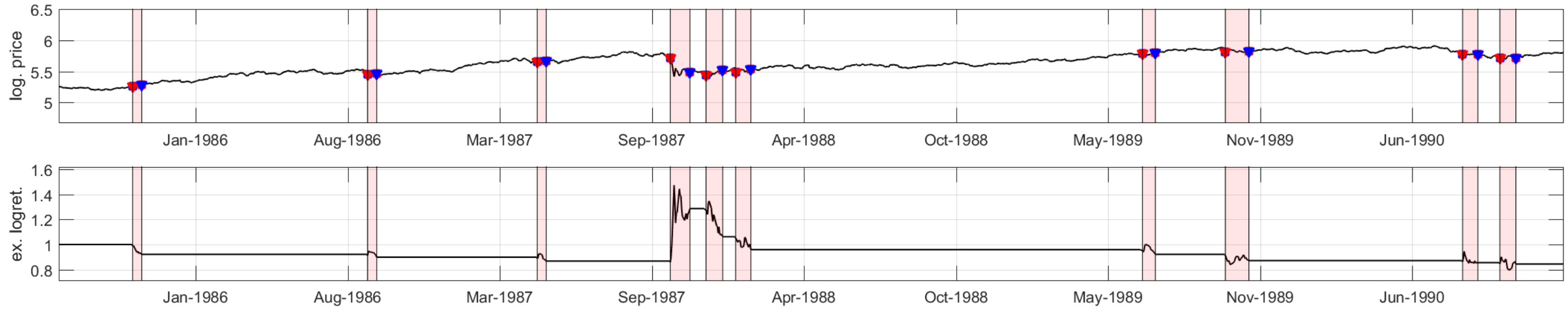
(a) January 1980 to December 2009



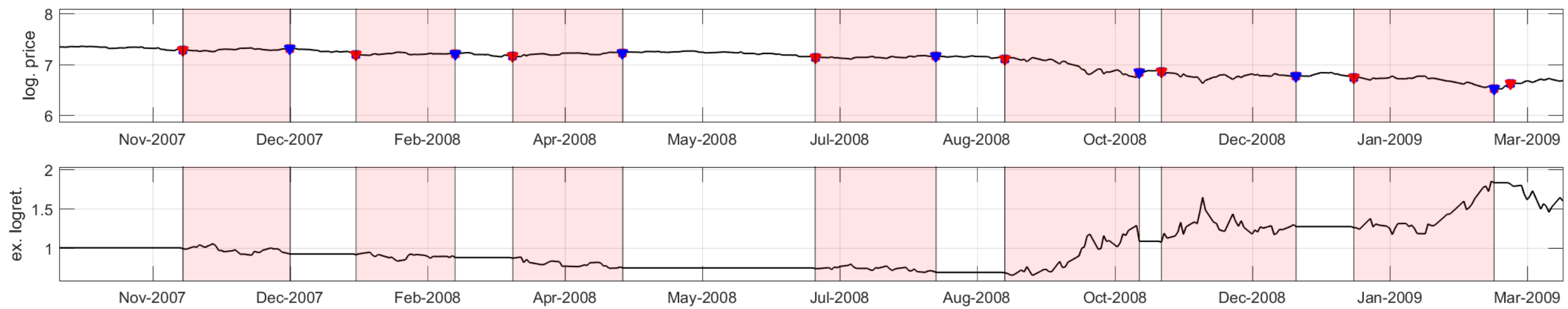
(b) January 1985 to December 1990

NOTES: Out-of-sample performance of the "LS"-entry-exit investment strategy a) between January 1980 and December 2009 b) between January 1985 and December 1990. Market regimes are estimated based on the data split with average period duration of 1 year.
Top panel: The S&P 500 index and signals from the entry-exit strategy, logarithmic scale. Highlighted areas are bear periods according to the entry-exit strategy.
Bottom panel: Cumulative returns on the exit-entry strategy in excess to returns on the buy-and-hold strategy, logarithmic scale.

Figure 4.2: Out-of-sample performance of the entry-exit investment strategy with T_{av} of 3 moths.



(a) January 1985 to December 1990



(b) October 2007 to March 2009

NOTES: Out-of-sample performance of the "LS"-entry-exit investment strategy a) between January 1985 and December 1990 b) between October 2007 and March 2009. Market regimes are estimated based on the data split with average period duration of 3 months.
Top panel: The S&P 500 index and signals from the entry-exit strategy, logarithmic scale. Highlighted areas are bear periods according to the entry-exit strategy.
Bottom panel: Cumulative returns on the exit-entry strategy in excess to returns on the buy-and-hold strategy, logarithmic scale.

Forró, Z. (2015). *Detecting bubbles in financial markets: Fundamental and dynamical approaches.* Ph. D. thesis, ETH Zurich.

Gapeev, P. V. (2010). Two switching multiple disorder problems for Brownian motions. Working paper.

Glabadanidis, P. (2014). The market timing power of moving averages: Evidence from US REITs and REIT indexes. *International Review of Finance 14*(2), 161–202.

Glabadanidis, P. (2015). Market timing with moving averages. *International Review of Finance 15*(3), 387–425.

Honda, T. (2003). Optimal portfolio choice for unobservable and regime-switching mean returns. *Journal of Economic Dynamics and Control 28*(1), 45–78.

Jiang, Z.-Q., W.-X. Zhou, D. Sornette, R. Woodard, K. Bastiaensen, and P. Cauwels (2010). Bubble diagnosis and prediction of the 2005–2007 and 2008–2009 Chinese stock market bubbles. *Journal of Economic Behavior and Organization 74*(3), 149–162.

Johansen, A., D. Sornette, and O. Ledoit (1999). Predicting financial crashes using discrete scale invariance. Working paper.

Koivu, M., T. Pennanen, and W. T. Ziemba (2005). Cointegration analysis of the FED model. *Finance Research Letters 2*(4), 248–259.

Li, C. and H. Xue (2009). A Bayesian's bubble. *The Journal of Finance 64*(6), 2665–2701.

Lleo, S. and W. T. Ziemba (2016). The bond-stock earnings yield differential model: Additional applications and other models for stock market crash prediction. *Quantitative Finance Letters 4*(1), 26–34.

Nguyen, D., J. Tie, and Q. Zhang (2014). An optimal trading rule under a switchable mean-reversion model. *Journal of Optimization Theory and Applications 161*(1), 145–163.

Phillips, P. C., Y. Wu, and J. Yu (2011). Explosive behavior in the 1990s Nasdaq: When did exuberance escalate asset values? *International Economic Review 52*(1), 201–226.

Shiryaev, A. and M. Zhitlukhin (2012a). Optimal stopping problems for a Brownian motion with a disorder on a finite interval. Working paper.

Shiryaev, A. N. (2002). Quickest detection problems in the technical analysis of the financial data. In *Mathematical Finance-Bachelier Congress 2000*, pp. 487–521. Springer.

Shiryaev, A. N. (2010). Quickest detection problems: Fifty years later. *Sequential Analysis 29*(4), 345–385.

Shiryaev, A. N., M. Zhitlukhin, and W. T. Ziemba (2014). When to sell Apple and the NASDAQ? Trading bubbles with a stochastic disorder model. *The Journal of Portfolio Management 40*(2), 54–63.

Shiryaev, A. N. and M. V. Zhitlukhin (2012b). Disorder detection problems with applications in finance. Working paper.

Shiryaev, A. N., M. V. Zhitlukhin, and W. T. Ziemba (2015). Land and stock bubbles, crashes and exit strategies in Japan circa 1990 and in 2013. *Quantitative Finance 15*(9), 1449–1469.

Sokko, A. (2015). Disorder problem for a Brownian motion on a segment in the case of uniformly distributed moment of disorder. *Moscow University Mathematics Bulletin 70*(3), 103–110.

Watanabe, K., H. Takayasu, and M. Takayasu (2007). A mathematical definition of the financial bubbles and crashes. *Physica A: Statistical Mechanics and its Applications 383*(1), 120–124.

Wu, Y. (1997). Rational bubbles in the stock market: Accounting for the US stock-price volatility. *Economic Inquiry 35*(2), 309.

Zhang, Q., D. Sornette, M. Balcilar, R. Gupta, Z. A. Ozdemir, and H. Yetkiner (2016). LPPLS bubble indicators over two centuries of the S&P 500 index. *Physica A: statistical Mechanics and its Applications 458*, 126–139.

Zhitlukhin, M. and A. Shiryaev (2013). Bayesian disorder problems on filtered probability spaces. *Theory of Probability and Its Applications 57*(3), 497–511.

Zhitlukhin, M. and W. Ziemba (2016). Exit strategies in bubble-like markets using a changepoint model. *Quantitative Finance Letters 4*(1), 47–52.

Zhitlukhin, M. V. (2014). *Stochastic dynamics of financial markets.* Ph. D. thesis, University of Manchester.

Ziemba, W., S. Lleo, and M. Zhitlukhin (2017). *Stock market crashes: Predictable and unpredictable and what to do about them*, Volume 13. World Scientific.

Appendix

A Posterior probability π_t and threshold function $b^*_\eta(t)$

The posterior probability π_t is given by

$$\pi_t = \frac{\Psi_t}{\Psi_t + 1 - G(t)}, \tag{A.1}$$

where Ψ_t is the Shiryaev-Roberts statistic, defined recurrently by

$$\begin{aligned} &\Psi_0 = 0, \\ &\Psi_t = \left(\frac{p}{T} + \Psi_{t-1}\right) \cdot \frac{\sigma_1}{\sigma_2} \cdot \exp\left\{\frac{(X_t-\mu_1)^2}{2\sigma_1^2} - \frac{(X_t-\mu_2)^2}{2\sigma_2^2}\right\}, \quad t = 1, \dots, T, \end{aligned}$$

with

$$G(t) = \mathbb{P}(\theta \le t), \quad p_t = \mathbb{P}(\theta = t).$$

The threshold function $b^*_\eta(t)$, $t = 0, \dots, T$, is given by

$$b^*_\eta(t) = \inf\big\{x \in [0,1] \mid V_\eta\big(t, (1-G(t)) \cdot x/(1-x)\big) = 0\big\} \tag{A.2}$$

with $V_\eta(t, y)$, $t = 0, \dots, T$, $y \in \mathbb{R}$, defined recursively by

$$\begin{aligned} &V_\eta(T, y) \equiv 0, \\ &V_\eta(t, x) = \begin{cases} \max\{0, \mu_2(x + p_{t+1}) + \mu_1(1 - G(t+1)) + f_1(t, x)\}, & \eta = 1, \\ \max\{0, (1-\eta)\beta^t[(\gamma-1)(x+p_{t+1}) & \\ +(\beta-1)(1-G(t+1))] + f_\eta(t,x)\}, & \eta \ne 1, \end{cases} \end{aligned}$$

where

$$\beta = \exp\left\{(1-\eta)\mu_1 + \frac{(1-\eta)^2\sigma_1^2}{2}\right\}, \quad \gamma = \exp\left\{(1-\eta)\mu_2 + \frac{(1-\eta)^2\sigma_2^2}{2}\right\}$$

and

$$\begin{aligned} f_\eta(t, x) = \int_{\mathbb{R}} V_\eta\left(t+1, (p_{t+1} + x) \cdot \frac{\sigma_1}{\sigma_2} \cdot \exp\left\{\frac{(z-\mu_1)^2}{2\sigma_1^2} - \frac{(z-\mu_2)^2}{2\sigma_2^2}\right\}\right) \\ \times \frac{1}{\sigma_1\sqrt{2\pi}} \exp\left\{-\frac{(z-\mu_1-(1-\eta)\sigma_1^2)^2}{2\sigma_1^2}\right\} dz. \end{aligned} \tag{A.3}$$

The threshold function $b^*_\eta(t)$ is decreasing in t.[15]

Computing the threshold function $b^*_\eta(t)$. To compute $b^*_\eta(t)$ one needs to recover functions $V(t,x)$ for every $t = 0,\dots,T-1$. Each function $V(t,\cdot)$ can be approximated by a piecewise constant function $\widetilde{V}(t,\cdot)$ with a step size ΔX on x.[16]

Obtaining functions $\widetilde{V}(t,x)$ is computationally very expensive, as integrals (A.3), or their centered versions $(z \to (z-\mu_1)/\sigma_1 - (1-\eta)\sigma_1)$,

$$f_\eta(t,x) = \int_{\mathbb{R}} \widetilde{V}_\eta\Bigg(t+1,(p_{t+1}+x)\cdot\frac{\sigma_1}{\sigma_2}\cdot\exp\Bigg\{\frac{(z+(1-\eta)\sigma_1)^2}{2} - \frac{(\sigma_1 z+(1-\eta)\sigma_1^2+\mu_1-\mu_2)^2}{2\sigma_2^2}\Bigg\}\Bigg) \times \frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}\,dz, \tag{A.4}$$

must be evaluated multiple times. The main difficulty arises from infinite integration limits in (A.4). Fixed truncation limits z^{min} and z^{max} lead to persistent underestimation of (A.4), with errors accumulating when $t \to 0$.[17] Increasing z^{min}, z^{max} in turn, requires more computational power.

The following method to compute the integral (A.4) allows to account for the tradeoff between precision and computation time. Without loss of generality we further assume that $\sigma_1 = \sigma_2 = \sigma$ and $\mu_1 > \mu_2$, hence only the positive infinity tail has to be approximated. The general case can be done analogously.

1. Select ΔZ (linear grid on z) and ε (integral precision);
2. Define
$$z^{min} = -\frac{\sigma}{(\mu_1-\mu_2)}\log\left(\frac{\overline{X}}{p_{t+1}+x}\right) - \sigma(1-\eta) - \frac{\mu_1-\mu_2}{2\sigma}$$
and
$$z^{max} = -\frac{\sigma}{(\mu_1-\mu_2)}\log\left(\frac{\Delta X}{p_{t+1}+x}\right) - \sigma(1-\eta) - \frac{\mu_1-\mu_2}{2\sigma},$$
where $\overline{X} = \inf\{x \geq 0 \mid \widetilde{V}(t+1,x) = 0\}$;
3. Define $z^{trunc} = z^{min} + \min(\Delta Z, z^{max} - z^{min})$;
4. Compute
$$\widetilde{f}_\eta(t,x) = \int_{z^{min}}^{z^{trunc}} \widetilde{V}_\eta(t+1,\dots) \times \frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}\,dz;$$

[15] Zhitlukhin (2014) shows that $V(t,x)$ is decreasing in both t and x. The latter gives

$$\begin{aligned} b^*_\eta(t+1) &= \inf\left\{x \in [0,1] \mid V_\eta\big(t+1,(1-G(t+1))\cdot x/(1-x)\big) = 0\right\} \\ &< \inf\left\{x \in [0,1] \mid V_\eta\big(t,(1-G(t+1))\cdot x/(1-x)\big) = 0\right\} \\ &< \inf\left\{x \in [0,1] \mid V_\eta\big(t,(1-G(t))\cdot x/(1-x)\big) = 0\right\} = b^*_\eta(t). \end{aligned}$$

[16] The choice ΔX of the linear grid step has been discussed in Zhitlukhin (2014).

[17] Let $g(z)$ be the mapping from z to the right argument of $\widetilde{V}$ in (A.4). If $\sigma_2 \geq \sigma_1$, then $\widetilde{V}(t+1,g(z))$ converges with $z\to\infty$ to $\max \widetilde{V}(t+1,\cdot) = \widetilde{V}(t+1,0)$.

5. Compute upper bound $E_\eta(z^{trunc})$ on the truncation error as

$$E_\eta(z^{trunc}) := \widetilde{V}(t+1,0)\cdot(\mathcal{N}(z^{max}) - \mathcal{N}(z^{trunc}))$$
$$\geq \int_{z^{trunc}}^{z^{max}} V_\eta(t+1,\dots)\times\frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}dz,$$

   where $\mathcal{N}(\cdot)$ is the CDF of the standard normal distribution;

6. If $E_\eta(z^{trunc}) < \varepsilon\cdot\widetilde{f}_\eta(t,x)$, approximate $f_\eta(t,x)$ by $\widetilde{f}_\eta(t,x)+\widetilde{V}(t+1,0)\cdot(1-\mathcal{N}(z^{trunc}))$; otherwise replace z^{trunc} by $z^{trunc}+\min(\Delta Z, z^{max}-z^{trunc})$ and repeat 2–4.

To keep integrals (A.4), and therefore truncation errors, of the same order across t it is recommended to work with normalized versions of $V(t,\cdot)$ (i.e., $V(t,0)\equiv 1$).

B Ex post identification of bull and bear periods in price data

The following procedure identifies (ex post) bull and bear periods in the risk factor's price data based on average duration T_{av} of periods.

The underlying idea is that the long term drift is positive in bull periods and negative in bear ones. A change of regime therefore occurs at some time between the moment when the risk factor exhibits a pronounced positive (negative) drift and the subsequent first moment when the long-term drift becomes significantly negative (positive). Small deviations of the drift from zero may be due to volatility rather than a change of regime.

1. Compute moving averages, $\mu_{ma,t}$, of 1-period logarithmic returns from $\frac{1}{4}T_{av}$ of past data (bottom panel in Figure B.1), remove observations with insufficient data;

2. Find the quantile

$$p(T_{av}) = \arg\min_{p\in(0,1)}|T_{av} - T(p)|,$$

   where

$$T(p) = \frac{1}{n}\sum_{i=1}^{n}(B_i(p) - A_i(p))$$

   is the average duration of bull/bear periods $[A_i(p), B_i(p)]$, $i = 1,\dots,n$, identified in the data based on quantile p (see 3. for periods' identification);

3. Split the data into bull/bear periods based on the quantile $p(T_{av})$:

- define $Th_{pos}(p)$ and $Th_{neg}(p)$ as quantiles of conditional empirical distributions of μ_{ma} with

$$P(\mu_{ma} < Th_{pos}(p) \mid \mu_{ma} > 0) = p$$

and

$$P(\mu_{ma} > Th_{neg}(p) \mid \mu_{ma} < 0) = p,$$

- define the hitting times

$$\begin{aligned} \Omega_{pos}^{hit}(p) &= \{t \mid \mu_{ma,t} \geq Th_{pos}(p) \text{ and } \mu_{ma,t-1} < Th_{pos}(p)\}, \\ \Omega_{pos}^{leave}(p) &= \{t \mid \mu_{ma,t} \geq Th_{pos}(p) \text{ and } \mu_{ma,t+1} < Th_{pos}(p)\}, \\ \Omega_{neg}^{hit}(p) &= \{t \mid \mu_{ma,t} \leq Th_{neg}(p) \text{ and } \mu_{ma,t-1} > Th_{neg}(p)\}, \end{aligned}$$

  * remove $t \in \Omega_{pos}^{hit}$ and $\theta(t) \in \Omega_{pos}^{leave}$, such that $\{s \in (\theta(t), t) \mid s \in \Omega_{neg}^{hit}(p)\} = \emptyset$ with $\theta(t) = \max\{s < t \mid s \in \Omega_{pos}^{leave}(p)\}$,
  * remove $t \in \Omega_{pos}^{leave}$ and $\theta(t) \in \Omega_{pos}^{hit}$, such that $\{s \in (t, \theta(t)) \mid s \in \Omega_{neg}^{hit}(p)\} = \emptyset$ with $\theta(t) = \min\{s > t \mid s \in \Omega_{pos}^{hit}(p)\}$,
- for each $t \in \Omega_{pos}^{hit}$ determine a *positive moving average period* $[a(t), b(t)] \in T_{ma}^{pos}(p)$ (blue areas in Figure B.1) with

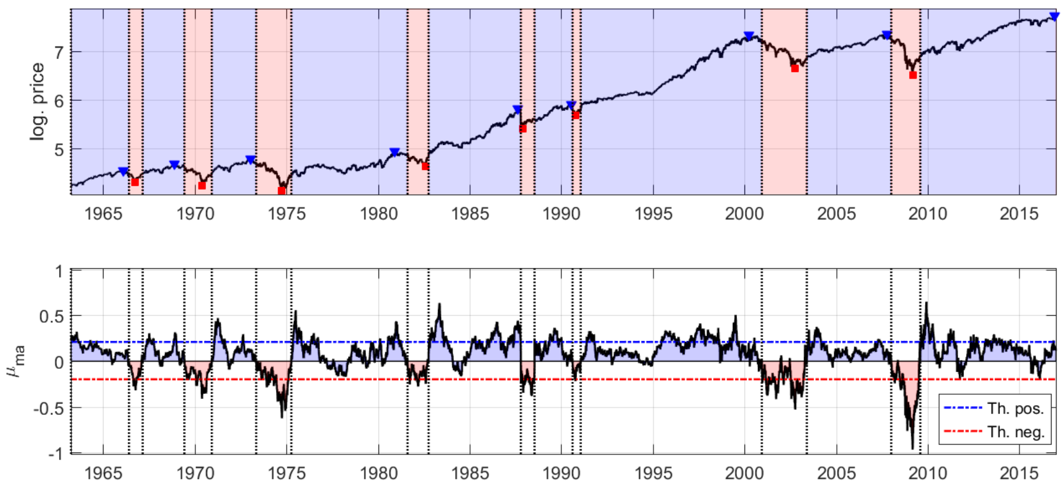
$$a(t) = \min\{s \leq t \mid \mu_{ma,l} \geq 0 \text{ for all } l \in [s, t]\}$$

and

$$b(t) = \begin{cases} \min\{s \in [t, \theta(t)] \mid \mu_{ma,l} \leq 0 \text{ for all } l \in [s, \theta(t)]\}, & \theta(t) \text{ defined,} \\ \max\{s \mid \mu_{ma,s} \text{ defined}\}, & \text{otherwise,} \end{cases}$$

with $\theta(t) = \min\{s > t \mid s \in \Omega_{neg}^{hit}\}$,
  * if $\xi := \min\{t \in \Omega_{pos}^{leave}\} < \min\{t \in \Omega_{pos}^{hit}\}$, let $T_{ma}^{pos}(p) = T_{ma}^{pos}(p) \cup [\min\{s \mid \mu_{ma,s} \text{ defined}\}, \xi]$,
- determine *negative moving average periods* as compliments to positive moving average periods (red areas in Figure B.1),
- define the set of *local price minimums (maximums)* as price minimums (maximums) within each negative (positive) moving average period (resp., red squares and blue triangles in Figure B.1),
- define each period between a local price minimum (maximum) and the following local price maximum (minimum) as a *bull (bear) period.*

Figure B.1: Identifying bull and bear periods in the S&P 500 price data.



NOTES: Top panel: The S&P 500 index and its local price maximums (red squares) and minimums (blue triangles) based on average period duration T_{av} of 1 year, logarithmic scale. Bottom panel: Annualized moving averages of 1-period logarithmic returns on the S&P 500 computed from past quarterly data.

Curriculum vitae

Anastasiia Sokko

Personal details

Date of birth:	22.10.1990
Nationality:	Russian

Education

09/2012 – 02/2018	Doctoral programme in Finance, University of Zurich and Swiss Finance Institute, Switzerland
09/2007 – 06/2012	Specialist in Mathematics (MSc in Mathematics), Lomonosov Moscow State University, Russia

Professional experience

09/2015 – 02/2018	Representative of academic assistants at the OEC Faculty Assembly, University of Zurich, Switzerland
09/2013 – 02/2018	Research and Teaching Assistant, Department of Banking and Finance, University of Zurich, Switzerland
06/2012 – 08/2012	Associate Intern in Quantitative Analytics, Barclays Capital, UK